

# A Tutorial on Kernel Density Estimation and Recent Advances

Yen-Chi Chen

Department of Statistics

University of Washington

April 14, 2017

*This tutorial provides a gentle introduction to kernel density estimation (KDE) and recent advances regarding confidence bands and geometric/topological features. We begin with a discussion of basic properties of KDE: the convergence rate under various metrics, density derivative estimation, and bandwidth selection. Then, we introduce common approaches to the construction of confidence intervals/bands, and we discuss how to handle bias. Next, we talk about recent advances in the inference of geometric and topological features of a density function using KDE. Finally, we illustrate how one can use KDE to estimate a cumulative distribution function and a receiver operating characteristic curve. We provide R implementations related to this tutorial at the end.*

**Keywords:** kernel density estimation, nonparametric statistics, confidence bands, bootstrap

## CONTENTS

1	Introduction . . . . .	2
2	Statistical Properties . . . . .	2
	2.1 Convergence Rate . . . . .	3
	2.2 Derivative Estimation . . . . .	6
	2.3 Bandwidth Selection . . . . .	6
3	Confidence Intervals and Confidence Bands . . . . .	7
	3.1 Confidence Intervals . . . . .	8
	3.1.1 Example: Confidence Intervals . . . . .	9
	3.2 Confidence Bands . . . . .	10
	3.2.1 Example: Confidence Bands . . . . .	11
	3.3 Handling the Bias . . . . .	12
	3.3.1 Ignoring the Bias . . . . .	12
	3.3.2 Undersmoothing . . . . .	13
	3.3.3 Bias-corrected and Oversmoothing . . . . .	13
4	Geometric and Topological Features . . . . .	14
	4.1 Local Modes . . . . .	15
	4.2 Level Sets . . . . .	16
	4.3 Ridges . . . . .	17
	4.4 Morse-Smale Complex . . . . .	17
	4.5 Cluster Trees . . . . .	18
	4.6 Persistent Diagram . . . . .	19
5	Estimating the CDF . . . . .	19
	5.1 ROC Curve . . . . .	20
6	Conclusion and Open Problems . . . . .	20
	Acknowledgement . . . . .	21
	References . . . . .	21

**1. Introduction.** Kernel density estimation (KDE), also known as the Parzen’s window (Parzen, 1962), is one of the most well-known approaches to estimate the underlying probability density function of a dataset. KDE is a nonparametric density estimator requiring no assumption that the underlying density function is from a parametric family. KDE will learn the shape of the density from the data automatically. This flexibility arising from its non-parametric nature makes KDE a very popular approach for data drawn from a complicated distribution.

Figure 1 illustrates KDE using a part of the NACC (National Alzheimers Coordinating Center) Uniform Data Set (Beekly et al., 2007), version 3.0 (March 2015). Because the purpose of using this dataset is to illustrate the effectiveness of KDE, we will draw no scientific conclusion but will just use KDE as a tool to explore the pattern of the data. We focus on two variables, ‘CRAFTDTI’ (Craft Story 21 Recall – delay time), and ‘CRAFTDVR’ (Craft Story 21 Recall – total story units recalled, verbatim scoring). Although these two variables take integer values, we treat them as continuous and use KDE to determine the density function. We consider only the unique subject with scores on both variables, resulting in a sample of size 4,044. In the left panel of Figure 1, we display the estimated density function of ‘CRAFTDTI’ using KDE. We see that there are two modes in the distribution. In the right panel of Figure 1, we show the scatter plot of the data and overlay it with the result of bivariate KDE (blue contours). Because many subjects have identical values for the two variables, the scatter plot (gray dots) provides no useful information regarding the underlying distribution. However, KDE shows the multi-modality of this bivariate distribution, which contains multiple bumps that cannot be captured easily by any parametric distribution.

The remainder of the tutorial is organized as follows. In Section 2, we present the definition of KDE, followed by a discussion of its basic properties: convergence rates, density derivative estimations, and bandwidth selection. Then, in Section 3, we introduce common approaches to the construction of confidence regions, and we discuss the problem of bias in statistical inference. Section 4 provides an introduction to the use of KDE to estimate geometric and topological features of a density function. In Section 5, we study how one can use KDE to estimate the cumulative distribution function (CDF) and the receiver operating characteristic (ROC) curve. Finally, in Section 6, we discuss open problems. At the end of this tutorial, we provide R codes for implementing the presented analysis of KDE.

**2. Statistical Properties.** Let  $X_1, \dots, X_n \in \mathbb{R}^d$  be an independent, identically distributed random sample from an unknown distribution  $P$  with density function  $p$ . Formally, KDE can be expressed as

$$(1) \quad \hat{p}_n(x) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right),$$

where  $K : \mathbb{R}^d \mapsto \mathbb{R}$  is a smooth function called the kernel function and  $h > 0$  is the smoothing bandwidth that controls the amount of smoothing. Two common examples of  $K(x)$  are

$$(\text{Gaussian kernel}) \quad K(x) = \frac{\exp(-\|x\|^2/2)}{v_{1,d}}, \quad v_{1,d} = \int \exp(-\|x\|^2/2) dx,$$

$$(\text{Spherical kernel}) \quad K(x) = \frac{I(\|x\| \leq 1)}{v_{2,d}}, \quad v_{2,d} = \int I(\|x\| \leq 1) dx.$$

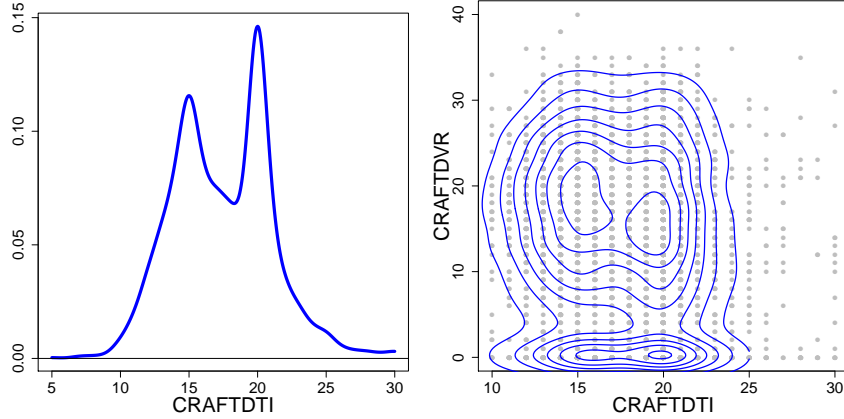


FIG 1. Examples of KDE using the NACC Uniform Data Set. We focus on variables ‘CRAFTDTI’ and ‘CRAFTDVR’ and use those subjects who have non-missing value of either variables. **Left:** We show the marginal density function of variable ‘CRAFTDTI’ from one-dimensional (1D) KDE. There are two bumps in for this density function. **Right:** We show the scatter plot along with the bivariate density function of both variables from the two-dimensional (2D) KDE, showing the multimodality feature of this bivariate density function.

Note that we apply the same amount of smoothing  $h$  in every direction; in practice, one can use a bandwidth matrix  $H$  and the quantity  $K\left(\frac{x-X_i}{h}\right)$  becomes  $K\left(H^{-1}(x-X_i)\right)$ .

Intuitively, KDE has the effect of smoothing out each data point into a smooth bump, whose the shape is determined by the kernel function  $K(x)$ . Then, KDE sums over all these bumps to obtain a density estimator. At regions with many observations, because there will be many bumps around, KDE will yield a large value. On the other hand, for regions with only a few observations, the density value from summing over the bumps is low, because only have a few bumps contribute to the density estimate.

Figure 2 presents examples of KDE in the 1D case. There are six observations, as indicated by the black lines. We smooth these observations into bumps (red bumps) and sum over all of the bumps to form the final density estimator (the blue curve). In R, many packages are equipped with programs for computing KDE; see [Deng and Wickham \(2011\)](#) for a listing.

**Remark.** (Adaptive smoothing) The amount of smoothing can depend on the location  $x$  ([Loftsgaarden and Quesenberry, 1965](#)) or the data point  $X_i$  ([Breiman et al., 1977](#)). In the former case, we use  $h = h(x)$  so KDE becomes  $\hat{p}_n(x) = \frac{1}{nh^d(x)} \sum_{i=1}^n K\left(\frac{x-X_i}{h(x)}\right)$ , which is referred to as the balloon estimator ([Terrell and Scott, 1992](#)). In the latter case, we use  $h = h_i = h(X_i)$  for the  $i$ -th data points with the resulting density estimate being  $\hat{p}_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_i^d} K\left(\frac{x-X_i}{h_i}\right)$ , which is referred to as the sample smoothing estimator ([Terrell and Scott, 1992](#)). For more details regarding adaptive smoothing, we refer the readers to Section 6.6 of [Scott \(2015\)](#).

**2.1. Convergence Rate.** To measure the errors of KDE, we consider three types of errors: the pointwise error, uniform error, and mean integrated square error (MISE). The pointwise error is the simplest error and is related to the confidence interval (Section 3.1). The uniform error has many useful theoretical properties since it measures the uniform deviation of the

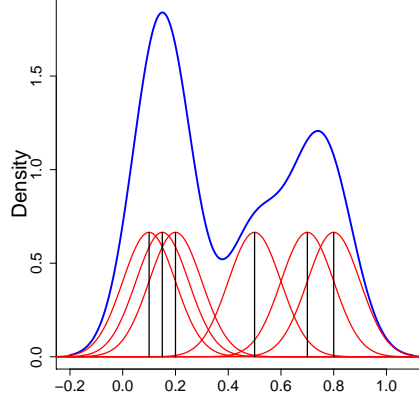


FIG 2. A 1D illustration of how KDE is constructed. There are six observations, located at the positions indicated by black lines. We then smooth these observations into small bumps (red bumps) and the sum them to obtain the density estimate (blue curve).

estimator and can be used to bound other types of errors. The uniform error is related to the confidence band and the geometric (and topological) features; see Section 3.2 and Section 4. The MISE (actually it is a risk measurement of the estimator) is generally used in bandwidth selection (Section 2.3), because it measures the overall performance of the estimator and is related to the mean square error.

**Pointwise Error.** For a given point  $x$ , the pointwise error of KDE is the difference between KDE  $\hat{p}_n(x)$  and  $p(x)$ , the true density function evaluated at  $x$ . Let  $\nabla^2 p = \sum_{\ell=1}^d \frac{\partial^2 p}{\partial x_\ell^2}$  be the Laplacian of the function  $p$ . Under smoothness conditions (Scott, 2015; Wasserman, 2006; Tsybakov, 2009),

$$\begin{aligned}
 \hat{p}_n(x) - p(x) &= \underbrace{\mathbb{E}(\hat{p}_n(x)) - p(x)}_{B_h(x)} + \underbrace{\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))}_{\mathcal{E}_n(x)} \\
 &= O(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right), \\
 B_h(x) &= \frac{h^2}{2} \sigma_K^2 \nabla^2 p(x) + o(h^2), \\
 \mathcal{E}_n(x) &= \sqrt{\frac{\mu_K \cdot p(x)}{nh^d}} \cdot Z_n(x) + o_P\left(\sqrt{\frac{1}{nh^d}}\right),
 \end{aligned}
 \tag{2}$$

where  $Z_n(x) \xrightarrow{D} N(0, 1)$  and  $\sigma_K^2 = \int \|x\|^2 K(x) dx$ ,  $\mu_K = \int K^2(x) dx$  are constants depending only on the kernel function  $K$ . Thus, when  $h \rightarrow 0$  and  $nh^d \rightarrow \infty$ ,  $\hat{p}_n(x) \xrightarrow{P} p(x)$ , i.e., KDE  $\hat{p}_n(x)$  is a consistent estimator of  $p(x)$ . Equation (2) presents the decomposition of the (pointwise) estimation error of KDE in terms of the bias  $B_h(x)$  and the stochastic variation  $\mathcal{E}_n(x)$ . This decomposition will be used frequently in deriving other errors and constructing the confidence regions.

**Uniform Error.** Another error metric is the uniform error (also known as the  $L_\infty$  error), the maximal difference between  $\hat{p}_n$  and  $p$ :  $\sup_x |\hat{p}_n(x) - p(x)|$ . According to empirical process theory (Yukich, 1985; Giné and Guillou, 2002; Einmahl and Mason, 2005; Rao, 2014), the uniform error

$$(3) \quad \sup_x |\hat{p}_n(x) - p(x)| = O(h^2) + O_P \left( \sqrt{\frac{\log n}{nh^d}} \right)$$

under mild conditions (see Giné and Guillou 2002; Einmahl and Mason 2005 for more details). The error rate in (3) and the pointwise error rate in (2) differ only in the stochastic variation portion and the difference is at the rate  $\sqrt{\log n}$ . The presence of an extra  $\sqrt{\log n}$  in the uniform error rate is a very common phenomenon in nonparametric estimation owing to empirical process theory. Generally, the uniform error is not practically useful, because it is very sensitive to small perturbations. However, the uniform error has many useful theoretical properties (Chen et al., 2015c; Chen, 2016; Fasy et al., 2014; Jisu et al., 2016), because it provides a uniform control of the estimation error over the entire support.

**MISE.** The MISE is one of the most well-known error measurements (Wasserman, 2006; Scott, 2015) among all of the error measures used in KDE. The MISE is defined as  $\int \mathbb{E} \left( (\hat{p}_n(x) - p(x))^2 \right) dx$ . Thus, the MISE measures the  $L_2$  risk of KDE. Under regularity conditions, the MISE

$$(4) \quad \begin{aligned} \int \mathbb{E} \left( (\hat{p}_n(x) - p(x))^2 \right) dx &= \int B_h^2(x) dx + \int \text{Var}(\mathcal{E}_n(x)) dx \\ &= \frac{h^4}{4} \sigma_K^2 \int \nabla^2 p(x) dx + \frac{\mu_K}{nh^d} + o(h^4) + o\left(\frac{1}{nh^d}\right). \end{aligned}$$

The MISE can be viewed as the mean square error of KDE. The dominating term  $\frac{h^4}{4} \sigma_K^2 \int \nabla^2 p(x) dx + \frac{\mu_K}{nh^d}$  is called the asymptotic mean integrated square error (AMISE). Equation (4) shows that the error (risk) of KDE can be decomposed into a bias component,  $\frac{h^4}{4} \sigma_K^2 \int \nabla^2 p(x) dx$ , and a variance component  $\frac{\mu_K}{nh^d}$  together with small corrections. This decomposition is known as the bias-variance tradeoff (Wasserman, 2006) and is very useful in practice because we can choose the smoothing bandwidth  $h$  by optimizing this error. If we ignore smaller order terms and use the AMISE, the minimal error occurs when we choose

$$(5) \quad h_{\text{opt}} = \left( \frac{4\mu_K}{\sigma_K^2 \int \nabla^2 p(x) dx} \cdot \frac{1}{n} \right)^{\frac{1}{d+4}}$$

which leads to the optimal MISE

$$\int \mathbb{E} \left( (\hat{p}_{n,\text{opt}}(x) - p(x))^2 \right) dx = \inf_{h>0} \int \mathbb{E} \left( (\hat{p}_n(x) - p(x))^2 \right) dx = O \left( n^{-\frac{2}{d+4}} \right).$$

Equation (5) will be a key result in choosing the smoothing bandwidth (Section 2.3). Note that in practice, people generally select the smoothing bandwidth by minimizing the MISE rather than other errors because (i) it is a risk function that does not depend on any particular sample, (ii) it measures the overall estimation error rather than putting too much weight on a small portion of the support (i.e., it is more robust to small perturbations), and (iii) it has

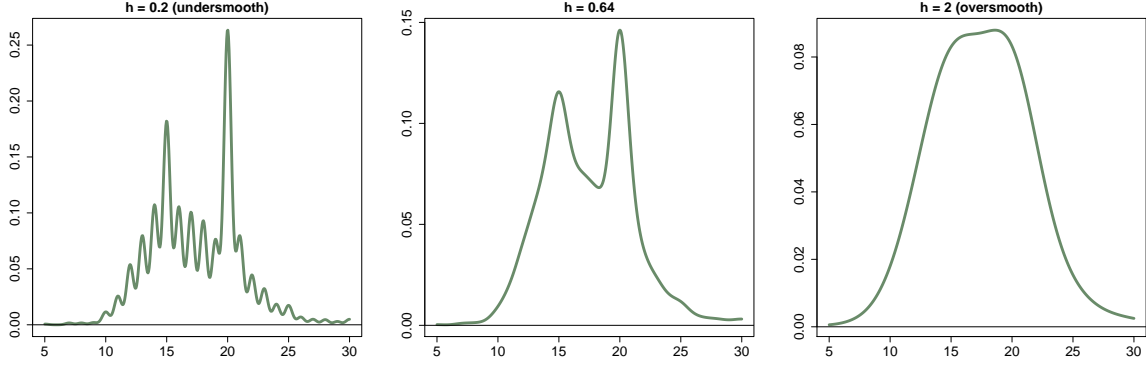


FIG 3. *Smoothing bandwidth and KDE.* We use the same data as in the left panel of Figure 1. We display KDE using three different amounts of smoothing. The left panel is the case of undersmoothing: we choose an excessively small bandwidth  $h$ . The middle panel is the case of the correct amount of smoothing, which is chosen according to the default rule in R. The right panel is the case of oversmoothing: the chosen  $h$  is too large.

useful theoretical behaviors, including the expression of the bias-variance tradeoff and the connection to the mean square error.

**Remark.** (Boundary bias) When the density function is discontinuous, the bias of KDE at the discontinuities will be of the order  $O(h)$  rather than  $O(h^2)$ , and this bias is called the boundary bias (Wasserman, 2006; Scott, 2015). In practice, one can use the boundary kernel to reduce the boundary bias (see, e.g., Chapter 6.2.3.5 in Scott 2015).

**2.2. Derivative Estimation.** KDE can be used to estimate the derivative of the density function. This is often called *density derivative estimation* (Stoker, 1993; Chacón et al., 2011). The idea is simple: we use the derivative of KDE as an estimator of the corresponding derivative of the density function. Let  $[\beta] = (\beta_1, \dots, \beta_d)$  be a multi-index (each  $\beta_\ell$  is a non-negative integer and  $||[\beta]|| = \sum_{\ell=1}^d \beta_\ell$ ). Define  $D^{[\beta]} = \frac{\partial^{\beta_1}}{\partial x_1^{\beta_1}} \cdots \frac{\partial^{\beta_d}}{\partial x_d^{\beta_d}}$  to be the  $[\beta]$ -th order partial derivative operator. Then, under smoothness assumptions (Chacón et al., 2011),

$$(6) \quad D^{[\beta]}\hat{p}_n(x) - D^{[\beta]}p(x) = O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+2||[\beta]||}}}\right).$$

That is, the (MISE or pointwise) error rate of gradients of KDE is  $O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+2}}}\right)$  and the error rate of second derivatives (Hessian matrix) of KDE is  $O(h^2) + O_P\left(\sqrt{\frac{1}{nh^{d+4}}}\right)$ . Similarly to  $B_h(x)$  and  $\mathcal{E}_n(x)$  in the density estimation, there are explicit formulas for the bias and stochastic variation of density derivative estimation; see Chacón et al. (2011) for more details. Some examples of using gradient and second derivative estimation can be found in Arias-Castro et al. (2016); Chacón and Duong (2013); Chen et al. (2015b,c); Genovese et al. (2009, 2014).

**2.3. Bandwidth Selection.** How to choose the smoothing bandwidth for KDE is a classical research topic in nonparametric statistics. This problem is often known as *bandwidth selection*.

Figure 3 shows KDE's with different amounts of smoothing of the same dataset. When  $h$  is too small (left panel), there are many wiggles in the density estimate. When  $h$  is too large (right panel), we smooth out important features. When  $h$  is at the correct amount (middle), we can see a clear picture of the underlying density.

Common approaches to bandwidth selection include the rule of thumb (Silverman, 1986), least square cross-validation, (Rudemo, 1982; Bowman, 1984; Bowman and Azzalini, 1997; Stone, 1984), biased cross-validation, (Scott and Terrell, 1987), and plug-in method (Woodroffe, 1970; Sheather and Jones, 1991). Roughly speaking, the core idea behind all of these methods is to minimize the AMISE, the dominating quantity in the MISE (4), or other similar error measurements. Different bandwidth selectors can be viewed as different estimators to the AMISE, and  $h$  is chosen by minimizing the AMISE estimator. Overviews and comparisons of the existing methods can be found in Jones et al. (1996); Sheather (2004), page 135–137 in Wasserman (2006), and Chapter 6.5 in Scott (2015).

While most of the literature focuses on the univariate case, Chacón and Duong (2013) provides a generalization of all of the above methods to the multivariate case and also generalizes the AMISE criterion into density derivative estimation. In R, one can use package ‘ks’<sup>1</sup> or ‘kedd’<sup>2</sup> to choose smoothing bandwidths for both density estimation and density derivative estimation. Note that the ks package is applicable to multivariate data.

In addition to the above approaches, Goldenshluger and Lepski (2011) propose a method, known as the Lepski’s approach (Goldenshluger and Lepski, 2008; Lepski and Goldenshluger, 2009), that treats the bandwidth selection problem as a model selection problem and proposes a new criterion for selecting the smoothing bandwidth. One feature of Lepski’s approach is that the selected bandwidth enjoys many statistical optimalities (Goldenshluger and Lepski, 2011).

**Remark.** (Kernel Selection) In contrast to bandwidth selection, the choice of kernel function does not play an important role in KDE. The effect of the kernel function on the estimation error is just a constant shift (via  $\sigma_K$  and  $\mu_K$  in equation (2)), and the difference is generally very small among common kernel functions (see, e.g., page 72 of Wasserman 2006 and Section 6.2.3 in Scott 2015), so most of the literature ignore this topic.

**3. Confidence Intervals and Confidence Bands.** Confidence regions of the density function are random intervals  $C_{1-\alpha}(x)$  derived from the sample such that  $C_{1-\alpha}(x)$  covers the true value of  $p(x)$  with probability at least  $1 - \alpha$ . Based on this notion, there are two common types of confidence regions:

- *Confidence interval*: for a given  $x$ , the set  $C_{1-\alpha}(x)$  satisfies

$$P(p(x) \in C_{1-\alpha}(x)) \geq 1 - \alpha.$$

- *Confidence band*: the interval  $C_{1-\alpha}(x)$  satisfies

$$P(p(x) \in C_{1-\alpha}(x) \forall x \in \mathbb{K}) \geq 1 - \alpha.$$

Namely, confidence intervals are confidence regions with only local coverage and confidence bands are confidence regions with simultaneous coverage. If a confidence interval/band has

<sup>1</sup><https://cran.r-project.org/web/packages/ks/index.html>

<sup>2</sup><https://cran.r-project.org/web/packages/kedd/index.html>



only coverage  $1 - \alpha + o_P(1)$ , it will be called an asymptotically valid  $1 - \alpha$  confidence interval/band.

For simplicity, we first ignore the bias between  $\hat{p}_n(x)$  and  $p(x)$  by assuming  $p(x) = \mathbb{E}(\hat{p}_n(x))$  in Sections 3.1 and 3.2 (i.e., we assume  $B_h(x) = 0$ ). We will discuss strategies for handling the bias in Section 3.3

3.1. *Confidence Intervals.* For a given point  $x$ , by equation (2),

$$(7) \quad \sqrt{nh^d}(\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))) = \sqrt{nh^d}\mathcal{E}_n \xrightarrow{d} N(0, \sigma_p^2(x)),$$

where  $\sigma_p^2(x) = \mu_K \cdot p(x)$ . Equation (7) implies that a straight-forward approach to construct a confidence band is to use asymptotic normality with a variance estimator.

**Method 1: Plug-in Approach.** A simple method is replacing  $p(x)$  in the asymptotic variance by its estimator  $\mu_K \hat{p}_n(x)$ , leading to the following  $1 - \alpha$  confidence interval of  $p(x)$ :

$$(8) \quad C_{1-\alpha, \text{PI}}(x) = \left[ \hat{p}_n(x) - z_{1-\alpha/2} \sqrt{\frac{\mu_K \cdot \hat{p}_n(x)}{nh^d}}, \quad \hat{p}_n(x) + z_{1-\alpha/2} \sqrt{\frac{\mu_K \cdot \hat{p}_n(x)}{nh^d}} \right].$$

We call this method the “plug-in method” because we plug-in the variance estimator to construct a confidence interval. When  $h \rightarrow 0, nh^d \rightarrow \infty$ ,  $\hat{p}_n(x)$  is a consistent estimator of  $p(x)$ . As a result,

$$P(\mathbb{E}(\hat{p}_n(x)) \in C_{1-\alpha, \text{PI}}(x)) = 1 - \alpha + o_P(1).$$

**Method 2: Bootstrap and Plug-in Approach.** An alternative method is to estimate the asymptotic variance using the bootstrap (Efron, 1979). In more detail, we use the empirical bootstrap (also known as the nonparametric bootstrap or Efron’s bootstrap, which is to sample the original data with replacement) to generate bootstrap sample  $X_1^*, \dots, X_n^*$ . Then, we apply KDE to the bootstrap sample, resulting in a bootstrap KDE  $\hat{p}_n^*(x)$ . When we repeat the bootstrap  $B$  times, we then have  $B$  bootstrap KDEs  $\hat{p}_n^{*(1)}(x), \dots, \hat{p}_n^{*(B)}(x)$ . Let

$$\hat{\sigma}_{p, \text{BT}}^2(x) = \frac{1}{B-1} \sum_{j=1}^B \left( \hat{p}_n^{*(j)}(x) - \bar{p}_n^*(x) \right)^2,$$

where  $\bar{p}_n^*(x) = \frac{1}{B} \sum_{j=1}^B \hat{p}_n^{*(j)}(x)$  is the sample average of the bootstrap KDE’s. Namely,  $\hat{\sigma}_{p, \text{BT}}^2(x)$  is the sample variance of the  $B$  bootstrap KDE’s evaluated at  $x$ . A bootstrap  $1 - \alpha$  confidence interval is

$$(9) \quad C_{1-\alpha, \text{BT+PI}}(x) = \left[ \hat{p}_n(x) - z_{1-\alpha/2} \cdot \hat{\sigma}_{p, \text{BT}}^2(x), \quad \hat{p}_n(x) + z_{1-\alpha/2} \cdot \hat{\sigma}_{p, \text{BT}}^2(x) \right].$$

Because the bootstrap variance estimator  $\hat{\sigma}_{p, \text{BT}}^2(x)$  converges to  $\frac{\sigma_p^2(x)}{nh^d}$  in the sense that

$$\frac{\hat{\sigma}_{p, \text{BT}}^2(x)}{\sigma_p^2(x)/(nh^d)} \xrightarrow{P} 1,$$

the bootstrap variance estimator is consistent, so the confidence interval will also be consistent:

$$P(\mathbb{E}(\hat{p}_n(x)) \in C_{1-\alpha, \text{BT+PI}}(x)) = 1 - \alpha + o_P(1).$$



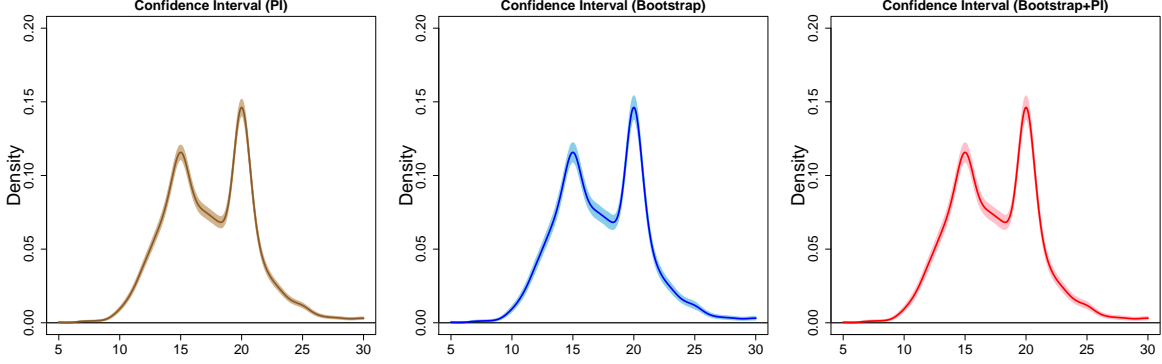


FIG 4. 95% confidence intervals from KDE. We use the same data as in the left panel of Figure 1. **Left:** We obtain confidence intervals using the plug-in approach (method 1 in Section 3.1). **Middle:** We construct confidence intervals using the plug-in approach with the bootstrap (method 2 in Section 3.1). **Right:** We build confidence intervals using the bootstrap approach (method 3 in Section 3.1). The three confidence regions are nearly the same, although they are constructed using different approaches. Note that there are two problems with these confidence regions. First, since we ignore the bias, the actual coverage might be substantially less than the nominal coverage 95%. Second, because they are confidence intervals, they only have pointwise coverage. Thus, even though the actual coverage is guaranteed, these regions might not cover the entire actual density function.

**Method 3: Bootstrap Approach.** In addition to the above methods, one can use a fully bootstrapping approach to construct a confidence interval without using asymptotic normality. Let  $\hat{p}_n^{*(1)}(x), \dots, \hat{p}_n^{*(B)}(x)$  be bootstrap KDE's as in the previous method. We define a pointwise deviation of a bootstrap KDE by

$$\Delta_1(x) = |\hat{p}_n^{*(1)}(x) - \hat{p}_n(x)|, \dots, \Delta_B(x) = |\hat{p}_n^{*(B)}(x) - \hat{p}_n(x)|.$$

Then we compute the  $1 - \alpha$  quantile of the empirical CDF of  $\Delta_1(x), \dots, \Delta_B(x)$ :

$$c_{1-\alpha, \text{BT}}(x) = \hat{G}_x^{-1}(1 - \alpha), \quad \hat{G}_x(t) = \frac{1}{B} \sum_{j=1}^B I(\Delta_j \leq t).$$

A  $1 - \alpha$  confidence interval of  $p(x)$  is

$$(10) \quad C_{1-\alpha, \text{BT}}(x) = [\hat{p}_n(x) - c_{1-\alpha, \text{BT}}(x), \quad \hat{p}_n(x) + c_{1-\alpha, \text{BT}}(x)].$$

Because the distribution of  $|\hat{p}_n^*(x) - \hat{p}_n(x)|$  approximates the distribution of  $|\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))|$ , this confidence interval is also asymptotically valid, i.e.,

$$P(\mathbb{E}(\hat{p}_n(x)) \in C_{1-\alpha, \text{BT}}(x)) = 1 - \alpha + o_P(1).$$

**3.1.1. Example: Confidence Intervals.** In Figure 4, we compare the three approaches of constructing a confidence interval using the NACC Uniform Data Set, as described in the Introduction (Section 1) and the left panel of Figure 1. The left panel is the 95% confidence interval of each point using the plug-in approach (method 1); the middle panel is the 95%

confidence interval from the plug-in and bootstrap approach (method 2); the right panel is the 95% confidence interval from the bootstrap approach (method 3).

Essentially, the three confidence intervals are very similar; in particular, the confidence intervals from the method 2 and 3 are nearly identical. The interval from method 1 is slightly smaller than the other two.

While all of these intervals are valid for each given point, there is no guarantee that they will cover the entire density function *simultaneously*. In the next section, we introduce methods of constructing confidence bands (confidence regions with simultaneous coverage).

**3.2. Confidence Bands.** Now, we present methods of constructing confidence bands. The key idea is to approximate the distribution of the uniform error  $\sup_x |\hat{p}_n(x) - p(x)|$  and then convert it into a confidence band. To be more specific, let  $G(t) = P(\sup_x |\hat{p}_n(x) - p(x)| < t)$  be the CDF of the uniform error, and let  $\bar{c}_{1-\alpha} = G^{-1}(1 - \alpha)$  be the  $1 - \alpha$  quantile. Then it can be shown that the set

$$\bar{C}(x) = [\hat{p}_n(x) - \bar{c}_{1-\alpha}, \hat{p}_n(x) + \bar{c}_{1-\alpha}]$$

is a confidence band, i.e.,

$$P(p(x) \in \bar{C}(x) \forall x \in \mathbb{K}) = 1 - \alpha.$$

Therefore, as long as we have a good approximation of the distribution  $G(t)$ , we can convert the approximation into a confidence band.

**Method 1: Plug-in Approach.** An intuitive approach is to derive the asymptotic distribution of  $\sup_x |\hat{p}_n(x) - p(x)|$  directly and then invert it into a confidence band. [Bickel and Rosenblatt \(1973\)](#); [Rosenblatt et al. \(1976\)](#) proved that the uniform loss converges to an extreme value distribution in the sense that

$$(11) \quad P\left(\sqrt{-2\log h} \left(\sqrt{nh^d} \sup_x \frac{|\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))|}{\sqrt{p(x)\mu_K}} - d_n\right) < t\right) \rightarrow e^{-2e^{-t}},$$

where  $d_n = O(\sqrt{-2\log h})$  is a quantity depending only on  $n, h$  and the kernel function  $K$ . [Rosenblatt et al. \(1976\)](#) provided an exact expression for the quantity  $d_n$ . Let  $E_{1-\alpha} = -\log\left(-\frac{\log \alpha}{2}\right)$  be the  $1 - \alpha$  quantile of the right-hand-side CDF. Define

$$c_{1-\alpha} = \sqrt{\frac{p(x)\mu_K}{nh^d}} \left(d_n + \frac{E_{1-\alpha}}{\sqrt{-2\log h}}\right).$$

Then, by equation (11),  $\sup_x |\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))|$  falls within  $[0, c_{1-\alpha}]$  with probability at least (asymptotically)  $1 - \alpha$ . To construct a confidence band, we replace the quantity  $p(x)$  in  $c_{1-\alpha}$  with a plug-in estimate from KDE, leading to

$$c_{1-\alpha, \text{PI}} = \sqrt{\frac{\hat{p}_n(x)\mu_K}{nh^d}} \left(d_n + \frac{E_{1-\alpha}}{\sqrt{-2\log h}}\right).$$

Then a  $1 - \alpha$  confidence band will be

$$(12) \quad C_{1-\alpha, \text{PI}}^\dagger(x) = [\hat{p}_n(x) - c_{1-\alpha, \text{PI}}, \hat{p}_n(x) + c_{1-\alpha, \text{PI}}].$$

Although equation (12) is an asymptotically valid confidence band, the convergence to the extreme value distribution in equation (11) is very slow. Thus, we need a huge sample size to guarantee that the confidence band from (12) is asymptotically valid. To resolve this problem, we use the bootstrap.

**Method 2: Bootstrap Approach.** The key element of how the bootstrap works is that the uniform error can be approximated accurately by the supremum of a Gaussian process (Neumann, 1998; Chernozhukov et al., 2014b,c). In more detail, there exists a tight Gaussian process  $B_n(x)$  such that

$$(13) \quad \sup_t \left| P \left( \sqrt{nh^d} \sup_x |\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))| < t \right) - P \left( \sup_x |B_n(x)| < t \right) \right| = o(1).$$

Moreover, the difference between the bootstrap KDE and the original KDE also has a similar convergent result (Chernozhukov et al., 2013, 2014a, 2016):

$$(14) \quad \sup_t \left| P \left( \sqrt{nh^d} \sup_x |\hat{p}_n^*(x) - \hat{p}_n(x)| < t | X_1, \dots, X_n \right) - P \left( \sup_x |B_n(x)| < t \right) \right| = o_P(1),$$

where  $B_n(x)$  is the same Gaussian process as the one in equation (13). Thus, the distribution of  $\sup_x |\hat{p}_n(x) - \mathbb{E}(\hat{p}_n(x))|$  will be approximated by the distribution of its bootstrap variety  $\sup_x |\hat{p}_n^*(x) - \hat{p}_n(x)|$ . As a result, the bootstrap quantile of uniform error converges to the quantile of the actual uniform error, thereby proving that the bootstrap confidence band is asymptotically valid.

Here is the formal construction of a bootstrap confidence band. Let  $\hat{p}_n^{*(1)}(x), \dots, \hat{p}_n^{*(B)}(x)$  be the bootstrap KDE's. We define the uniform deviation of the bootstrap KDE's by

$$\Delta_1 = \sup_x |\hat{p}_n^{*(1)}(x) - \hat{p}_n(x)|, \dots, \Delta_B = \sup_x |\hat{p}_n^{*(B)}(x) - \hat{p}_n(x)|.$$

Then, we compute the  $1 - \alpha$  quantile of the empirical CDF of  $\Delta_1, \dots, \Delta_B$  as

$$c_{1-\alpha, \text{BT}} = \hat{G}_{\mathbb{K}}^{-1}(1 - \alpha), \quad \hat{G}_{\mathbb{K}}(t) = \frac{1}{B} \sum_{j=1}^B I(\Delta_j(x) \leq t).$$

A  $1 - \alpha$  confidence band will be

$$(15) \quad C_{1-\alpha, \text{BT}}^\dagger(x) = [\hat{p}_n(x) - c_{1-\alpha, \text{BT}}, \quad \hat{p}_n(x) + c_{1-\alpha, \text{BT}}].$$

By equations (13) and (14), when  $B \rightarrow \infty$ ,

$$P \left( \mathbb{E}(\hat{p}_n(x)) \in C_{1-\alpha, \text{BT}}^\dagger(x) \quad \forall x \in \mathbb{K} \right) = 1 - \alpha + o_P(1).$$

Namely, the set  $C_{1-\alpha, \text{BT}}^\dagger(x)$  is an asymptotically valid  $1 - \alpha$  confidence band.

**3.2.1. Example: Confidence Bands.** Figure 5 presents confidence bands using KDE. The left panel shows the confidence band from the bootstrap approach introduced in the previous section. Compared to the confidence intervals in Figure 4, the confidence bands are wider because we need to control the coverage simultaneously for every point.

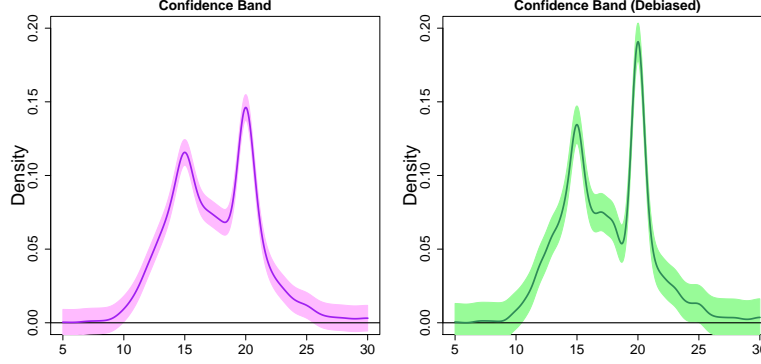


FIG 5. 95% confidence bands from KDE via the bootstrap. This is the same dataset as in the left panel of Figure 1. **Left:** The confidence band from bootstrapping the uniform error of KDE (method 2 of Section 3.2). Although this is a confidence band, we ignore the bias in constructing the confidence band so the actual coverage could be below the nominal coverage. **Right:** The confidence band from bootstrapping the uniform error of the debiased KDE (method proposed in Figure 6). Because we are using the debiased KDE, the density curve is slightly different from the blue curve in the left panel. Although the confidence band is wider than the confidence in the left, this confidence band has coverage guaranteed when the smoothing bandwidth is chosen at rate  $O\left(n^{-\frac{1}{d+4}}\right)$ . Note that there are possible approach to narrowing down the size of this confidence band; we refer the reader to [Chen \(2017\)](#).

However, the confidence band in the left panel of Figure 5 (and all of the confidence intervals in Figure 4) has a serious problem—the coverage guarantee is for the expected value of KDE  $\mathbb{E}(\hat{p}_n(x))$  rather than for the true density function  $p$ . Unless we undersmooth KDE (choose  $h$  converging at a faster rate than  $O(n^{-\frac{1}{d+4}})$ ), the confidence band shows undercoverage. We will discuss this topic in more detail in Section 3.3.

To construct an (asymptotically valid) confidence band with  $h$  being at rate  $O(n^{-\frac{1}{d+4}})$ , we use the debiased estimator introduced in [Chen \(2017\)](#). The details are provided in Section 3.3.3 and Figure 6. The right panel of Figure 5 shows a confidence band from the debiased KDE approach. Although the confidence band is wider, the coverage is guaranteed for such a confidence band.

**3.3. Handling the Bias.** In the previous section, we ignored the bias in KDE. However, the bias could be a severe problem in reality because it systematically shifted our confidence interval/band so the actual coverage is below the nominal coverage. Here we discuss strategies to handle bias.

**3.3.1. Ignoring the Bias.** A simple strategy is to ignore the bias and focus on inferring the expectation of KDE  $p_h(x) = \mathbb{E}(\hat{p}_n(x))$ .  $p_h$  is called the smoothed or mollified density function ([Rinaldo and Wasserman, 2010](#); [Chen et al., 2015b,c](#)). As long as the kernel function  $K$  is smooth,  $p_h$  will also be smooth. Moreover,  $p_h$  exists even when the distribution function is singular (in this case, the population density  $p$  does not exist). For inferring geometric or topological features,  $p_h$  might be a better parameter of interest because structures in  $p_h$  generally represent salient structures of  $p$  ([Chen et al., 2015c](#)) and many topological structures of  $p_h$  will be similar to those of  $p$  when  $h$  is small ([Chen et al., 2015b](#); [Genovese et al., 2014](#);

[Jisu et al., 2016](#)). If we switch our target to  $p_h$ , we have to make it clear that this is a confidence region of  $p_h$  rather than of  $p$  when we report our confidence regions.

**3.3.2. Undersmoothing.** Undersmoothing is a very common approach to handle bias in KDE (and other nonparametric approaches). Recall from equation (2):

$$\begin{aligned}\widehat{p}_n(x) - p(x) &= B_h(x) = \mathcal{E}_n(x) \\ &= O(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right).\end{aligned}$$

The bias is  $B_h(x) = O(h^2)$ , and the stochastic variation is the  $\mathcal{E}_n(x) = O_P\left(\sqrt{\frac{1}{nh^d}}\right)$  term. If now we take  $h \rightarrow 0$  such that  $h^2 = o\left(\sqrt{\frac{1}{nh^d}}\right)$ , then the stochastic variability dominates the errors. Therefore, we can ignore the bias term and use the method suggested in Section 3.1 and 3.2.

Note that  $h^2 = o\left(\sqrt{\frac{1}{nh^d}}\right)$  is equivalent to  $nh^{d+4} \rightarrow 0$ , which corresponds to choosing a smaller smoothing bandwidth than the optimal smoothing bandwidth ( $h_{\text{opt}} = O(n^{-\frac{1}{d+4}})$ ) that balances the bias and the variance (this is why it is called undersmoothing). Although the undersmoothing provides a valid construction of confidence regions, such a choice of bandwidth implies that the size of the confidence band is larger than the optimal size because we are inflating the variance to eliminate the bias. Some references to undersmoothing can be found in [Bjerve et al. \(1985\)](#); [Hall \(1992a\)](#); [Hall and Owen \(1993\)](#); [Chen \(1996\)](#); [Neumann and Polzehl \(1998\)](#); [Chen and Qin \(2002\)](#); [McMurry and Politis \(2008\)](#).

**3.3.3. Bias-corrected and Oversmoothing.** An alternative approach to construct a valid confidence band is to correct the bias of KDE explicitly; this approach is known as the bias-corrected method and the resulting KDE is called the bias-corrected KDE. Recall from equation (2) that the bias in KDE is

$$\mathbb{E}(\widehat{p}_n(x)) - p(x) = \frac{h^2}{2} \sigma_K^2 \nabla^2 p(x) + o(h^2).$$

Thus, we can correct the bias by estimating  $\nabla^2 p(x)$ . The quantity  $\nabla^2 p(x)$  can be estimated by  $\nabla^2 \widehat{p}_b(x)$ , where

$$\widehat{p}_b(x) = \frac{1}{nb^d} \sum_{i=1}^n K\left(\frac{x - X_i}{b}\right)$$

is KDE using smoothing bandwidth  $b$ . Recall from Section 2.2 that the second derivative estimator has an error rate

$$(16) \quad \nabla^2 \widehat{p}_b(x) - \nabla^2 p(x) = O(b^2) + O_P\left(\sqrt{\frac{1}{nb^{d+4}}}\right).$$

Thus, to obtain a consistent estimator of  $\nabla^2 p(x)$ , we have to choose another smoothing bandwidth  $b$ , and this smoothing bandwidth needs to be larger than the optimal bandwidth

$h_{\text{opt}} = O(n^{-\frac{1}{d+4}})$ . Because the choice of  $b$  corresponds to oversmoothing KDE, this approach is also called the “oversmoothing” method.

Using  $\hat{p}_b(x)$ , the bias-corrected KDE is

$$\tilde{p}_n(x) = \hat{p}_n(x) - \frac{h^2}{2} \sigma_K^2 \nabla^2 \hat{p}_b(x).$$

When  $\nabla^2 \hat{p}_b(x)$  is a consistent estimator of  $\nabla^2 p(x)$ , the pointwise error rate is

$$\tilde{p}_n(x) - p(x) = o(h^2) + o_P(h^2) + O_P\left(\sqrt{\frac{1}{nh^d}}\right),$$

so the dominating quantity,  $O_P\left(\sqrt{\frac{1}{nh^d}}\right)$ , is the stochastic variation in  $\tilde{p}_n(x)$ . Thus, the confidence regions can be constructed by replacing  $\hat{p}_n(x)$  by  $\tilde{p}_n(x)$  in equations (8), (9), (10), and (15). An incomplete list of literature of the bias-corrected approach is as follows: [Härdle and Bowman \(1988\)](#); [Härdle and Marron \(1991\)](#); [Hall \(1992b\)](#); [Eubank and Speckman \(1993\)](#); [Sun and Loader \(1994\)](#); [Härdle et al. \(1995\)](#); [Neumann \(1995\)](#); [Xia \(1998\)](#); [Härdle et al. \(2004\)](#).

[Calonico et al. \(2015\)](#) proposes a plug-in method of constructing a confidence interval with  $b = h$ . Although the choice  $b = h$  does not lead to a consistent estimate of the second derivative, the bias will be pushed into the next order because the bias-corrected estimator can be viewed as a higher order kernel function (see page 157 in [Scott 2015](#)). Thus, since the dominating term in the estimation error is the stochastic variation, we can construct an asymptotically valid confidence interval using the plug-in approach.

[Chen \(2017\)](#) further generalizes this idea to confidence bands by bootstrapping the bias-corrected kernel density estimator with  $b = h$ . Figure 6 summarizes the procedure of constructing a confidence band using the approach in [Chen \(2017\)](#). The resulting confidence band,  $\tilde{C}_{1-\alpha, \text{BT}}^\dagger(x)$ , has the following property:

$$P\left(p(x) \in \tilde{C}_{1-\alpha, \text{BT}}^\dagger(x) \ \forall x \in \mathbb{K}\right) = 1 - \alpha + o_P(1)$$

when  $h = h_{\text{opt}} = O(n^{-\frac{1}{d+4}})$ . Namely,  $\tilde{C}_{1-\alpha, \text{BT}}^\dagger(x)$  is an asymptotically valid  $1 - \alpha$  confidence band when we pick  $h$  under the optimal rate. The right panel of Figure 5 shows an example of this confidence band. Although this approach generally leads to a wider confidence band, this confidence band has asymptotically  $1 - \alpha$  coverage whereas the confidence band in the left panel of Figure 5 has undercoverage.

**Remark.** (Calibration) In addition to the above methods, another possible approach is to choose a corrected coverage of confidence regions; this approach is called “calibration” and is related to the work in [Beran \(1987\)](#); [Hall \(1986\)](#); [Loh \(1987\)](#); [Hall and Horowitz \(2013\)](#). The principal idea is to investigate the effect of the bias on the coverage of the confidence band and then choose a conservation quantile to guarantee the nominal coverage of the resulting confidence regions.

**4. Geometric and Topological Features.** KDE can be used to estimate not only the underlying density function but also geometric (and topological) structures related to the density. To be more precise, many geometric (and topological) features of  $\hat{p}_n$  converges to the

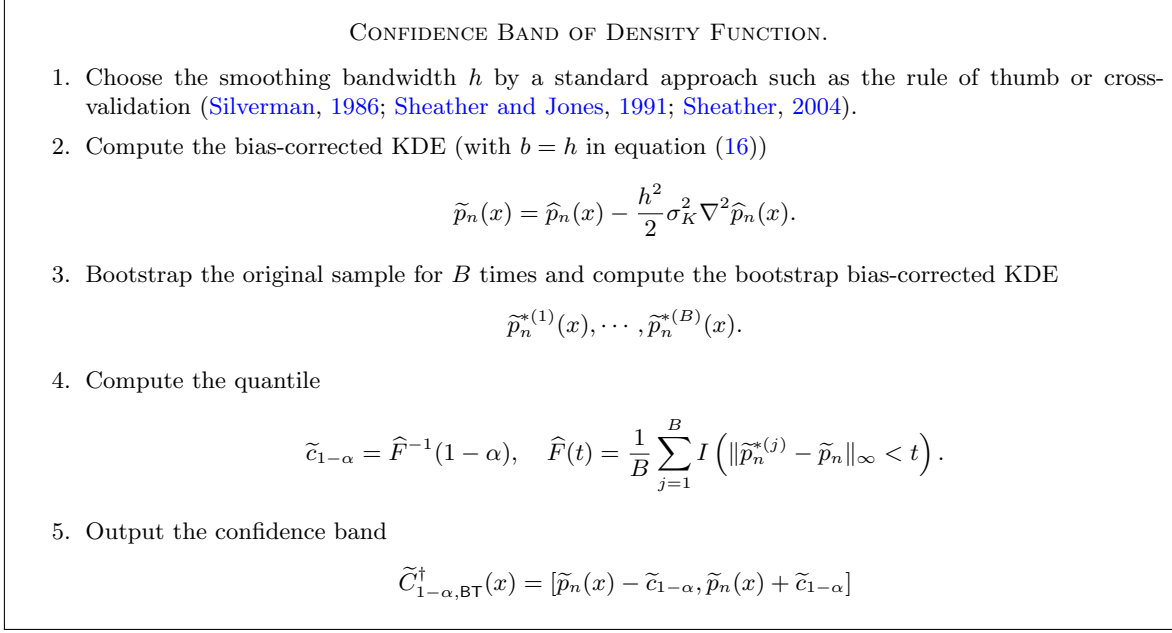


FIG 6. A confidence band of the density function from bootstrapping the debiased KDE (Chen, 2017). This confidence band is asymptotically valid and is compatible with most of the bandwidth selectors introduced in Section 2.3.

corresponding structures of  $p$ , and hence, we can use a structure of  $\hat{p}_n$  as the estimator of that structure of  $p$ .

Because geometric and topological structures generally involve the gradient and Hessian matrix of the density function, we define some notations here. We define  $g(x) = \nabla p(x)$  to be the gradient of the density and  $H(x) = \nabla \nabla p(x)$  to be the Hessian matrix of the density. Moreover, we also define  $\lambda_1(x) \geq \dots \geq \lambda_d(x)$  to be the largest to the smallest eigenvalues of  $H(x)$  and  $v_1(x), \dots, v_d(x)$  to be the corresponding eigenvectors.

**4.1. Local Modes.** A well-known geometric feature of the density function is its (global) mode. Actually, when Parzen introduced KDE, he mentioned the use of the mode of KDE to estimate the mode of the density function (Parzen, 1962). The asymptotic distribution and confidence sets of the mode were later discussed in Romano (1988a,b).

We can extend the definition of the (global) mode to a local sense and define the local modes:

$$\mathcal{M} = \{x : g(x) = 0, \lambda_1(x) < 0\}.$$

Namely,  $\mathcal{M}$  is the collection of points for which the density function is locally maximized. A natural estimator of  $\mathcal{M}$  is a plug-in from KDE (Chazal et al., 2014; Chen et al., 2016):

$$\widehat{\mathcal{M}} = \{x : \hat{g}_n(x) = 0, \hat{\lambda}_1(x) < 0\},$$

where  $\hat{g}_n(x)$  and  $\hat{\lambda}_1(x)$  are KDE version of  $g(x)$  and  $\lambda_1(x)$ . Under mild assumptions,  $\widehat{\mathcal{M}}$  is a consistent estimator of  $\mathcal{M}$  (Chazal et al., 2014; Chen et al., 2016). Note that one can use



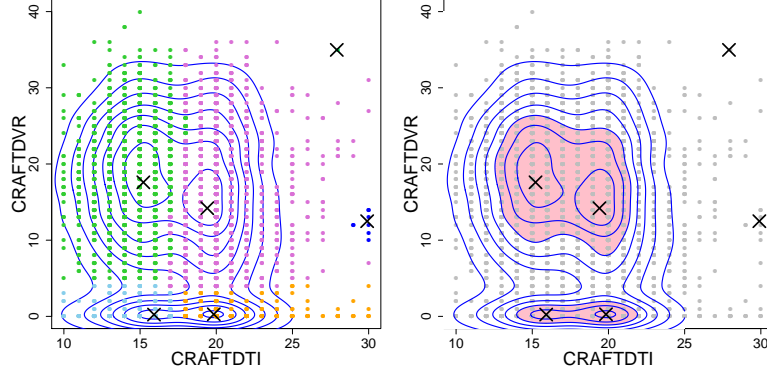


FIG 7. Estimating geometric features using KDE. This is the same dataset as the right panel of Figure 1. **Left:** Density local modes (black crosses) and mode clustering. The colored points describe the clusters that the respective subject belong to. Mode clustering uses the gradient flow to partition data points into clusters. Thus, each cluster (points with the same color) has a local mode as its representative. **Right:** Density contours (blue), local modes (black crosses), and a density level set (pink area). A density level set is just a region containing points whose density values are greater than or equal to a particular level. Thus, it will contain regions within some contour lines (blue curves).

the mean shift algorithm (Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu and Meer, 2002) to compute the estimator  $\widehat{\mathcal{M}}$  numerically.

Note that one can use the local modes to cluster data points; this is called mode clustering (Chacón et al., 2015; Azizyan et al., 2015; Chen et al., 2016) or mean-shift clustering (Fukunaga and Hostetler, 1975; Cheng, 1995; Comaniciu and Meer, 2002). The left panel of Figure 7 shows a case of estimated local modes (black boxes) and mode clustering using the mean-shift algorithm. In R, one can use the library ‘LPCM<sup>3</sup>’ to compute the estimator  $\widehat{\mathcal{M}}$  and perform mode clustering.

**4.2. Level Sets.** Level sets are regions for which the density value is equal to or above a particular level. Given a level  $\lambda$ , the  $\lambda$ -level set (Polonik, 1995; Tsybakov, 1997) is

$$L_\lambda = \{x : p(x) \geq \lambda\}.$$

A natural estimator of  $L_\lambda$  is the plug-in estimate from KDE:

$$\widehat{L}_\lambda = \{x : \widehat{p}_n(x) \geq \lambda\}.$$

The pink area in the right panel of Figure 7 is one instance of the density level set.

There is substantial statistical literature discussing different types of convergence of  $\widehat{L}_\lambda$ ; see Polonik (1995); Tsybakov (1997); Rinaldo and Wasserman (2010); Rinaldo et al. (2012) and the references therein. Mammen and Polonik (2013) and Chen et al. (2015c) propose procedures for constructing confidence sets of  $L_\lambda$  through bootstrapping the estimator  $\widehat{L}_\lambda$ . Note that a visualization tool<sup>4</sup> for multivariate level sets is proposed in Chen et al. (2015c).

<sup>3</sup><https://cran.r-project.org/web/packages/LPCM/index.html>

<sup>4</sup>R source code: <https://github.com/yenchic/HDLV>

4.3. *Ridges.* Another interesting geometric structures are ridges (Genovese et al., 2014; Chen et al., 2015b; Qiao et al., 2016) of the density functions. Formally, ridges are defined as follows. Let  $V(x) = [v_2(x) \cdots v_d(x)] \in \mathbb{R}^{d \times (d-1)}$  be the matrix consisting of the second eigenvector to the last eigenvector. A density ridge is defined as

$$\mathcal{R} = \{x : V(x)V(x)^T g(x) = 0, \lambda_2(x) < 0\}.$$

Intuitively, any point  $x \in \mathcal{R}$  is a local mode in the subspace spanned by  $v_2(x), \dots, v_d(x)$ . Thus, if we move away from  $\mathcal{R}$  in the subspace, the density value decreases, which is the characteristic attribute of a ridge.

To estimate  $\mathcal{R}$ , we again use the plug-in from KDE:

$$\hat{\mathcal{R}} = \left\{x : \hat{V}(x)\hat{V}(x)^T \hat{g}(x) = 0, \hat{\lambda}_2(x) < 0\right\},$$

where  $\hat{V}(x)$  and  $\hat{\lambda}_2(x)$  are KDE versions of  $V(x)$  and  $\lambda_2(x)$  respectively. The convergence rate and topological characteristics were discussed in Genovese et al. (2014). Chen et al. (2015b) and Qiao et al. (2016) both studied the asymptotic theory, and Chen et al. (2015b) further proposed methods of constructing confidence sets of  $\mathcal{R}$ . Ozertem and Erdogmus (2011) introduced the subspace-constrained mean shift (SCMS) algorithm to compute  $\hat{\mathcal{R}}$ . The red curves in the right panel of Figure 7 are estimated ridges from the SCMS algorithm.

4.4. *Morse-Smale Complex.* The Morse-Smale complex (Banyaga, 2004) of a density function  $p$  is a partition of the entire support  $\mathbb{K}$  based on the density gradient flow. For any point  $x \in \mathbb{K}$ , we define a gradient ascent flow  $\pi_x(t)$  such that

$$\pi'_x(t) = g(\pi_x(t)), \quad \pi_x(0) = 0.$$

Namely,  $\pi_x(t)$  is a flow starting at  $x$  such that we move along the orientation of the density gradient ascent. By Morse theory (Morse, 1925, 1930), such a flow converges to a destination that is one of the critical points (points where  $g(x) = 0$ ) when the density function is smooth. The colored curves in the left panel of Figure 7 are gradient ascent flows  $\pi_x(t)$  of KDE starting at each data point.

Similarly, we define a gradient descent flow  $\gamma_x(t)$  such that

$$\gamma'_x(t) = -g(\gamma_x(t)), \quad \gamma_x(0) = 0.$$

In a similar manner to  $\pi_x(t)$ ,  $\gamma_x(t)$  starts at  $x$  but now the flow moves by following density gradient descent. Again, by Morse theory, such a flow converges also to one of the critical points, but not to the same point as the destination of  $\pi_x(t)$ . Based on the destinations of  $\pi_x(t)$  and  $\gamma_x(t)$ , we partition the entire support  $\mathbb{K}$  into different regions; points within the same region share the same destination for both the gradient ascent flow and the gradient descent flow. This partition is called the Morse-Smale complex.

To estimate the Morse-Smale complex of  $p$ , we use the Morse-Smale complex of  $\hat{p}_n$ . Arias-Castro et al. (2016) and Chen et al. (2015d) studied the convergence of gradient flows and the Morse-Smale complex of  $\hat{p}_n$  and proved the statistical consistency of these geometric features. Chen et al. (2015d) further proposed to use the Morse-Smale complex to visualize a

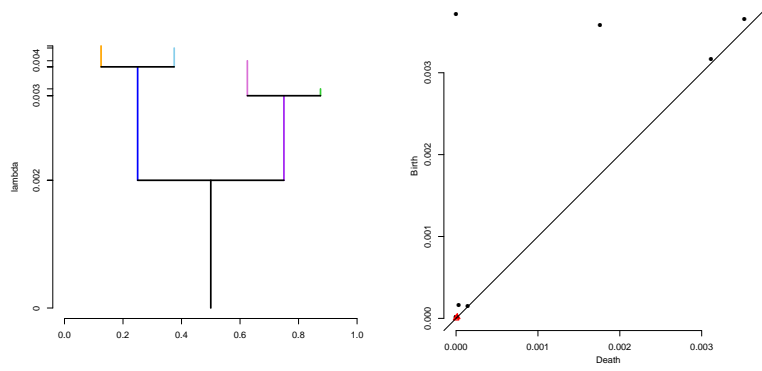


FIG 8. Estimating topological features of KDE. This is the same dataset as in the right panel of Figure 1. **Left:** Cluster tree, a tree structure representing KDE. The four leaves correspond to the four high-density local modes in Figure 7 (other local modes are tiny so the cluster tree algorithm ignores them). **Right:** The persistent diagram. The top four black dots denote the four persistent topological features (four connected components) of KDE, which are created by the four high-density local modes in Figure 7. There are several structures in the bottom left corner; they correspond to the topological noises in constructing a persistent diagram. For more details, we refer the reader to Wasserman (2016).

multivariate density function<sup>5</sup>. Note that one can use the R package ‘msr’<sup>6</sup> that to perform data analysis with the Morse-Smale complex (see Gerber et al. 2010; Gerber and Potter 2011; Gerber et al. 2013 for more details).

**4.5. Cluster Trees.** Cluster trees (also known as density trees Stuetzle 2003; Klemelä 2009) are tree-structured objects summarizing the structure of the underlying density function. The left panel of Figure 8 provides an example of a cluster tree corresponding to KDE in the right panel of Figure 2 (also the same dataset as in Figure 7 and 8). A cluster tree is constructed as follows. Recall that  $L_\lambda = \{x : p(x) \geq \lambda\}$  is the density level set at the level  $\lambda$ . When the level  $\lambda$  is too large,  $L_\lambda$  will be empty, because no region has a density value above such level. When we gradually decrease  $\lambda$  from a large number, at some levels (when  $\lambda$  hits the density value of local modes), a new connected component will be created; this corresponds the creation of a connected component. Moreover, at particular levels, two or more connected components will merge (generally at the density value of local minima or saddle points); this corresponds to the elimination of a connected component. Cluster tree uses a tree structure to summarize the creation and elimination of connected components at different levels. Since a cluster tree always live in a 2D plane, it is an excellent tool for visualizing a multivariate density function. For more details, we refer to the review paper in Wasserman (2016).

The above defines a cluster tree using the underlying population density  $p$ . In practice, we can construct an estimated cluster tree using KDE  $\hat{p}_n$ . Convergence of the cluster tree estimator was studied in Balakrishnan et al. (2013); Chen (2016). Jisu et al. (2016) provides a procedure for constructing confidence sets of cluster trees based on bootstrapping KDE  $\hat{p}_n$ . In R, one can use the package ‘TDA’<sup>7</sup> to construct a tree estimator of KDE.

<sup>5</sup>R source code: [https://github.com/yenchic/Morse\\_Smale](https://github.com/yenchic/Morse_Smale)

<sup>6</sup><https://cran.r-project.org/web/packages/msr/index.html>

<sup>7</sup><https://cran.r-project.org/web/packages/TDA/index.html>

**4.6. Persistent Diagram.** A persistent diagram (Cohen-Steiner et al., 2007; Edelsbrunner and Morozov, 2012; Wasserman, 2016) is a diagram summarizing the topological features of a density function  $p$ . The construction of a persistent diagram is very similar to that of a cluster tree, but now we focus on not only the connected components but also higher-order topological structures, such as loops and voids (see Fasy et al. 2014; Wasserman 2016 for a more details).

There are several means of estimating a persistent diagram. A natural approach is to use the persistent diagram of KDE  $\hat{p}_n$ . For such an estimator, the stability theorem in Cohen-Steiner et al. (2007) together with the uniform convergence in Section 2.1 are sufficient to prove the convergence of the estimated persistent diagram toward the population persistent diagram. For statistical inference, Fasy et al. (2014) proposes a bootstrap approach over KDE to construct a confidence set. In practice, one can use R package ‘TDA’<sup>8</sup> to construct the persistent diagram of KDE.

The right panel of Figure 8 shows an example of the persistent diagram of connected components (zeroth-order topological features) of KDE described in the right panel of Figure 2. At the top of the figure, the four dots indicate the existence of the four high-density local modes. In the bottom left regions, the black dots and red triangles are the topological noises representing low-density local modes (black dots) and low-density “loops” structures (red triangles; first-order topological features). Note that the two low-density local modes (black crosses) in the right part of both panels of Figure 7 are topological noises corresponding to the two black dots in the bottom left corner of the persistent diagram.

**5. Estimating the CDF.** KDE can also be used to estimate the CDFs (Nadaraya, 1964). The estimator is simple; we just integrate KDE (for simplicity, here we consider the univariate case):

$$(17) \quad \hat{F}_n(x) = \int_{-\infty}^x \hat{p}_n(y) dy.$$

Convergence of such estimators was extensively analyzed soon after their introduction (Winter, 1973; Azzalini, 1981; Reiss, 1981; Fernholz, 1991; Yukich, 1992; Mack, 1984).

Similarly to the pointwise error rate, one can show that

$$\hat{F}_n(x) - F(x) = O(h^2) + O_P\left(\sqrt{\frac{1}{n}}\right) + O_P\left(\sqrt{\frac{h}{n}}\right)$$

(see the derivation in Nadaraya 1964 and Azzalini 1981). Again, the first quantity  $O(h^2)$  is related to the bias. The other two quantities are related to the stochastic variation. Under such a rate, the optimal smoothing bandwidth will be  $h^* = O(n^{-1/3})$ , which leads to the error rate

$$\hat{F}_n(x) - F(x) = O_P\left(\sqrt{\frac{1}{n}}\right),$$

the same as using the empirical CDF. Note that as long as  $h = O(n^{-1/4})$ , we will obtain the square root error rate.

---

<sup>8</sup><https://cran.r-project.org/web/packages/TDA/index.html>

To construct a confidence band of  $F(x)$  via  $\hat{F}_n(x)$ , one can use the uniform central limit theorem proposed by [Giné and Nickl \(2008\)](#) to relate the uniform loss to the supremum of a Gaussian process and then use either the limiting distribution or the bootstrap. Note that to apply the result in [Giné and Nickl \(2008\)](#), one have to undersmooth (see Theorem 4 and Section 4.1.1 in [Giné and Nickl 2008](#)) the data or use a higher order kernel function (see Remark 7 and Corollary 2 in [Giné and Nickl 2008](#)).

**5.1. ROC Curve.** KDE can also be applied to estimate and infer the ROC curve ([McNeil and Hanley, 1984](#)). In the setting of the ROC curve, we observed two samples. The first is the sample of healthy subjects, whose responses are  $X_1, \dots, X_n$  from an unknown density  $P$ . The other is the sample of diseased individuals, whose response are  $Y_1, \dots, Y_m$  from an unknown density  $G$ . Consider a simple rule of classification based on choosing a cutoff point  $s$  such that we classify an individual as diseased if its response value is larger than  $s$ , otherwise it is classified as a healthy individual.

For such a rule, the sensitivity is  $SE(s) = 1 - G(s)$ , the probability of detecting a diseased subject. We also define the specificity  $SP(s) = F(s)$ , the probability of successfully assigning a healthy subject to the healthy group. Then, the ROC curve is defined as the plotting of the true positive fraction  $SE(s)$  versus the false positive fraction  $1 - SP(s)$ , or equivalently, as plotting the function

$$ROC(t) = 1 - G(F^{-1}(1 - t)).$$

A recent review on ROC curves can be found in [Demidenko \(2012\)](#).

A classical nonparametric approach of estimating  $ROC(t)$  is to plug-in the empirical CDF estimator for both  $F$  and  $G$  ([Hsieh et al., 1996](#)). As an alternative, one can use the integrated KDEs of both samples to estimate the ROC curve ([Zou et al., 1997](#); [Zhou and Harezlak, 2002](#); [Hall et al., 2004](#)) as equation (17). This is often called a smoothed estimator of the ROC curve because the resulting ROC curve estimator is generally a smooth curve.

To construct confidence bands of an ROC curve, most methods propose using the plug-in estimate from the empirical distribution and constructing the confidence band by the bootstrap ([Moise et al., 1985](#); [Campbell, 1994](#); [Macaskassy and Provost, 2004](#)). A formal proof of the theoretical validity of such a bootstrap approach is provided in [Hall et al. \(2004\)](#); [Horváth et al. \(2008\)](#); [Bertail et al. \(2009\)](#). Note that can also construct a confidence band by bootstrapping the smoothed ROC curve estimator and using the method proposed in Section 3.2 to construct a confidence band.

**6. Conclusion and Open Problems.** In this tutorial, we reviewed KDE's basic properties and its applications in estimating structures of the underlying density function. For readers who would like to learn more about different varieties of KDE, we recommend [Wasserman \(2006\)](#) and [Scott \(2015\)](#). Because this is a tutorial, we ignore many advanced topics such as the minimax theory and adaptation. An introduction of these theoretical properties can be found in [Tsybakov \(2009\)](#).

Although KDE has been widely studied since its introduction in the 1960s, there are still open problems that deserve further investigation. Here we briefly discuss some open problems related to the materials in this tutorial.

- **Confidence bands of other KDE-type estimators.** In addition to estimating a probability density function and its related structures, the idea of kernel smoothing can

be applied to estimate a regression, hazard, or survival function. Moreover, in casual inference, we might be interested in the difference between the regression/hazard/survival function from the control group and that from the treatment group as a characteristic of the treatment effect. One example is the conditional average treatment effect (Lee and Whang, 2009; Hsu, 2013; Ma and Zhou, 2014; Abrevaya et al., 2015). Although in this tutorial we have seen methods of constructing confidence bands of density functions, how to construct a (asymptotically) valid confidence band of these functions remains an open question.

- **Multidimensional problems.** When the dimension of the data  $d$  is large, KDE poses several challenges. First, KDE (and most other nonparametric density estimators) suffers severely from the so-called *curse of dimensionality*: The optimal convergence rate  $O(n^{-\frac{2}{d+4}})$  is very slow when  $d$  is large, and this slow convergence rate cannot be improved (Stone, 1982; Tsybakov, 2009) unless we assume extra smoothness. One way to solve this problem is to find density surrogates that can be estimated easily and to switch our parameter of interest to a density surrogate. However, this rises the question of what the correct surrogates and the corresponding estimators are, and this still remains unclear. Another issue of KDE in multi-dimensions is visualization. When  $d > 3$ , we can no longer see the entire KDE, and we must therefore use visualization tools to explore our density estimates. However, it is still unclear how to choose a visualization tool in practice.
- **More about geometric/topological structures.** In Section 4, we saw that several useful geometric and topological structures can be estimated by the corresponding structures of KDE. However, we do not yet fully understand the behavior of these estimators. For instance, how to choose the smoothing bandwidth that optimally estimate these structures is unclear. Handling this issue may require generalizing the concept of the MISE to the set estimator (Chen et al., 2015a) and choosing the smoothing bandwidth that minimizes such an error measurement. In addition to bandwidth selection, uniform inference remains an open question for these structures. Although there are methods of constructing confidence sets of most of these structures, it is unclear whether the resulting confidence sets are uniform for a collection of density functions. Moreover, theoretical optimality, such as the minimax theory, remains unclear for several of these structures, presenting another set of open questions in the study of KDE.

**Acknowledgement.** We thank Gang Chen and Larry Wasserman for useful comments and suggestions.

## References.

- J. Abrevaya, Y.-C. Hsu, and R. P. Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.
- E. Arias-Castro, D. Mason, and B. Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, 17(43):1–28, 2016.
- M. Azizyan, Y.-C. Chen, A. Singh, and L. Wasserman. Risk bounds for mode clustering. *arXiv preprint arXiv:1505.00482*, 2015.
- A. Azzalini. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, pages 326–328, 1981.
- S. Balakrishnan, S. Narayanan, A. Rinaldo, A. Singh, and L. Wasserman. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems*, pages 2679–2687, 2013.

- A. Banyaga. *Lectures on Morse homology*, volume 29. Springer Science & Business Media, 2004.
- D. L. Beekly, E. M. Ramos, W. W. Lee, W. D. Deitrich, M. E. Jacka, J. Wu, J. L. Hubbard, T. D. Koepsell, J. C. Morris, W. A. Kukull, et al. The national alzheimer’s coordinating center (nacc) database: the uniform data set. *Alzheimer Disease & Associated Disorders*, 21(3):249–258, 2007.
- R. Beran. Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3):457–468, 1987.
- P. Bertail, S. J. Cléménçon, and N. Vayatis. On bootstrapping the roc curve. In *Advances in Neural Information Processing Systems*, pages 137–144, 2009.
- P. Bickel and M. Rosenblatt. On some global measures of the deviations of density function estimates. *The Annals of Statistics*, 1(6):1071–1095, 1973.
- S. Bjerre, K. A. Doksum, and B. S. Yandell. Uniform confidence bounds for regression based on a simple moving average. *Scandinavian Journal of Statistics*, pages 159–169, 1985.
- A. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.
- A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, 1997.
- L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of multivariate densities. *Technometrics*, 19(2):135–144, 1977.
- S. Calonico, M. D. Cattaneo, and M. H. Farrell. On the effect of bias estimation on coverage accuracy in nonparametric inference. *arXiv preprint arXiv:1508.02973*, 2015.
- G. Campbell. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in medicine*, 13(5-7):499–508, 1994.
- J. E. Chacón and T. Duong. Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electronic Journal of Statistics*, 7:499–532, 2013.
- J. E. Chacón, T. Duong, and M. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, 21:807–840, 2011.
- J. E. Chacón et al. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015.
- F. Chazal, B. T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Robust topological inference: Distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197*, 2014.
- S. X. Chen. Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika*, 83(2):329–341, 1996.
- S. X. Chen and Y. S. Qin. Confidence intervals based on local linear smoother. *Scandinavian Journal of Statistics*, 29(1):89–99, 2002.
- Y.-C. Chen. Generalized cluster trees and singular measures. *arXiv preprint arXiv:1611.02762*, 2016.
- Y.-C. Chen. Nonparametric inference via bootstrapping the debiased estimator. *arXiv preprint arXiv:1702.07027*, 2017.
- Y.-C. Chen, C. R. Genovese, S. Ho, and L. Wasserman. Optimal ridge detection using coverage risk. In *Advances in Neural Information Processing Systems*, pages 316–324, 2015a.
- Y.-C. Chen, C. R. Genovese, and L. Wasserman. Asymptotic theory for density ridges. *The Annals of Statistics*, 43(5):1896–1928, 2015b.
- Y.-C. Chen, C. R. Genovese, and L. Wasserman. Density level sets: Asymptotics, inference, and visualization. *arXiv:1504.05438*, 2015c.
- Y.-C. Chen, C. R. Genovese, and L. Wasserman. Statistical inference using the morse-smale complex. *arXiv preprint arXiv:1506.08826*, 2015d.
- Y.-C. Chen, C. R. Genovese, and L. Wasserman. A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, 10(1):210–241, 2016.
- Y. Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790–799, 1995.
- V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013.
- V. Chernozhukov, D. Chetverikov, and K. Kato. Anti-concentration and honest, adaptive confidence bands. *The Annals of Statistics*, 42(5):1787–1818, 2014a.
- V. Chernozhukov, D. Chetverikov, and K. Kato. Comparison and anti-concentration bounds for maxima of gaussian random vectors. *Probability Theory and Related Fields*, pages 1–24, 2014b.
- V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximation of suprema of empirical processes.



- The Annals of Statistics*, 42(4):1564–1597, 2014c.
- V. Chernozhukov, D. Chetverikov, and K. Kato. Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related gaussian couplings. *Stochastic Processes and their Applications*, 2016.
- D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.
- D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002.
- E. Demidenko. Confidence intervals and bands for the binormal roc curve revisited. *Journal of applied statistics*, 39(1):67–79, 2012.
- H. Deng and H. Wickham. Density estimation in r. <http://vita.had.co.nz/papers/density-estimation.pdf>, 2011.
- H. Edelsbrunner and D. Morozov. Persistent homology: theory and practice. In *Proceedings of the European Congress of Mathematics*, pages 31–50, 2012.
- B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.
- U. Einmahl and D. M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005.
- R. L. Eubank and P. L. Speckman. Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, 88(424):1287–1301, 1993.
- B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.
- L. T. Fernholz. Almost sure convergence of smoothed empirical distribution functions. *Scandinavian Journal of Statistics*, pages 255–262, 1991.
- K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32–40, 1975.
- C. R. Genovese, M. Perone-Pacifco, I. Verdinelli, and L. Wasserman. On the path density of a gradient field. *The Annals of Statistics*, 37(6A):3236–3271, 2009.
- C. R. Genovese, M. Perone-Pacifco, I. Verdinelli, and L. Wasserman. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511–1545, 2014.
- S. Gerber and K. Potter. Data analysis with the morse-smale complex: The msr package for r. *Journal of Statistical Software*, 2011.
- S. Gerber, P.-T. Bremer, V. Pascucci, and R. Whitaker. Visual exploration of high dimensional scalar functions. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1271–1280, 2010.
- S. Gerber, O. Rübel, P.-T. Bremer, V. Pascucci, and R. T. Whitaker. Morse-smale regression. *Journal of Computational and Graphical Statistics*, 22(1):193–214, 2013.
- E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l’Institut Henri Poincaré (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier, 2002.
- E. Giné and R. Nickl. Uniform central limit theorems for kernel density estimators. *Probability Theory and Related Fields*, 141(3-4):333–387, 2008.
- A. Goldenshluger and O. Lepski. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 2008.
- A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3):1608–1632, 2011.
- P. Hall. On the bootstrap and confidence intervals. *The Annals of Statistics*, 14(4):1431–1452, 1986.
- P. Hall. On bootstrap confidence intervals in nonparametric regression. *The Annals of Statistics*, 20(2):695–711, 1992a.
- P. Hall. Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *The Annals of Statistics*, 20(2):675–694, 1992b.
- P. Hall and J. Horowitz. A simple bootstrap method for constructing nonparametric confidence bands for functions. *The Annals of Statistics*, 41(4):1892–1921, 2013.
- P. Hall and A. B. Owen. Empirical likelihood confidence bands in density estimation. *Journal of Computational and Graphical Statistics*, 2(3):273–289, 1993.
- P. Hall, R. J. Hyndman, and Y. Fan. Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*, 91(3):743–750, 2004.
- W. Härdle and A. W. Bowman. Bootstrapping in nonparametric regression: local adaptive smoothing and

- confidence bands. *Journal of the American Statistical Association*, 83(401):102–110, 1988.
- W. Härdle and J. Marron. Bootstrap simultaneous error bars for nonparametric regression. *The Annals of Statistics*, 19(2):778–796, 1991.
- W. Härdle, S. Huet, and E. Jolivet. Better bootstrap confidence intervals for regression curve estimation. *Statistics: A Journal of Theoretical and Applied Statistics*, 26(4):287–306, 1995.
- W. Härdle, S. Huet, E. Mammen, and S. Sperlich. Bootstrap inference in semiparametric generalized additive models. *Econometric Theory*, 20(02):265–300, 2004.
- L. Horváth, Z. Horváth, and W. Zhou. Confidence bands for roc curves. *Journal of Statistical Planning and Inference*, 138(6):1894–1904, 2008.
- F. Hsieh, B. W. Turnbull, et al. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The annals of statistics*, 24(1):25–40, 1996.
- Y.-C. Hsu. Consistent tests for conditional treatment effects. Technical report, Institute of Economics, Academia Sinica, Taipei, Taiwan, 2013.
- K. Jisu, Y.-C. Chen, S. Balakrishnan, A. Rinaldo, and L. Wasserman. Statistical inference for cluster trees. In *Advances In Neural Information Processing Systems*, pages 1831–1839, 2016.
- M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996.
- J. Klemelä. *Smoothing of multivariate data: density estimation and visualization*, volume 737. John Wiley & Sons, 2009.
- S. Lee and Y.-J. Whang. Nonparametric tests of conditional treatment effects. 2009.
- O. Lepski and A. Goldenshluger. Structural adaptation via lp-norm oracle inequalities. *Probab. Theory Related Fields*, 126(1-2):47–71, 2009.
- D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.
- W.-Y. Loh. Calibrating confidence coefficients. *Journal of the American Statistical Association*, 82(397):155–162, 1987.
- Y. Ma and X.-H. Zhou. Treatment selection in a randomized clinical trial via covariate-specific treatment effect curves. *Statistical methods in medical research*, page 0962280214541724, 2014.
- Y. Mack. Remarks on some smoothed empirical distribution functions and process. *Bulletin of informatics and cybernetics*, 21(1/2):29–35, 1984.
- S. Macskassy and F. Provost. Confidence bands for roc curves: Methods and an empirical study. Proceedings of the First Workshop on ROC Analysis in AI. August 2004., 2004.
- E. Mammen and W. Polonik. Confidence regions for level sets. *Journal of Multivariate Analysis*, 122:202–214, 2013.
- T. L. McMurry and D. N. Politis. Bootstrap confidence intervals in nonparametric regression with built-in bias correction. *Statistics & Probability Letters*, 78(15):2463–2469, 2008.
- B. J. McNeil and J. A. Hanley. Statistical approaches to the analysis of receiver operating characteristic (roc) curves. *Medical decision making*, 4(2):137–150, 1984.
- A. Moise, B. Clément, P. Ducimetière, and M. G. Bourassa. Comparison of receiver operating curves derived from the same population: a bootstrapping approach. *Computers and biomedical research*, 18(2):125–131, 1985.
- M. Morse. Relations between the critical points of a real function of  $n$  independent variables. *Transactions of the American Mathematical Society*, 27(3):345–396, 1925.
- M. Morse. The foundations of a theory of the calculus of variations in the large in  $m$ -space (second paper). *Transactions of the American Mathematical Society*, 32(4):599–631, 1930.
- E. A. Nadaraya. Some new estimates for distribution functions. *Theory of Probability & Its Applications*, 9(3):497–500, 1964.
- M. H. Neumann. Automatic bandwidth choice and confidence intervals in nonparametric regression. *The Annals of Statistics*, 23(6):1937–1959, 1995.
- M. H. Neumann. Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. *Annals of Statistics*, pages 2014–2048, 1998.
- M. H. Neumann and J. Polzehl. Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics*, 9(4):307–333, 1998.
- U. Ozertem and D. Erdogmus. Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 2011.

- E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.
- W. Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics*, pages 855–881, 1995.
- W. Qiao, W. Polonik, et al. Theoretical analysis of nonparametric filament estimation. *The Annals of Statistics*, 44(3):1269–1297, 2016.
- B. P. Rao. *Nonparametric functional estimation*. Academic press, 2014.
- R.-D. Reiss. Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, pages 116–119, 1981.
- A. Rinaldo and L. Wasserman. Generalized density clustering. *The Annals of Statistics*, 38(5):2678–2722, 2010.
- A. Rinaldo, A. Singh, R. Nugent, and L. Wasserman. Stability of density-based clustering. *The Journal of Machine Learning Research*, 13(1):905–948, 2012.
- J. P. Romano. Bootstrapping the mode. *Annals of the Institute of Statistical Mathematics*, 40(3):565–586, 1988a.
- J. P. Romano. On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics*, pages 629–647, 1988b.
- M. Rosenblatt et al. On the maximal deviation of  $k$ -dimensional density estimates. *The Annals of Probability*, 4(6):1009–1015, 1976.
- M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pages 65–78, 1982.
- D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.
- D. W. Scott and G. R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, 82(400):1131–1146, 1987.
- S. Sheather and C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53(3):683–690, 1991.
- S. J. Sheather. Density estimation. *Statistical Science*, 19(4):588–597, 2004.
- B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.
- T. M. Stoker. Smoothing bias in density derivative estimation. *Journal of the American Statistical Association*, 88(423):855–863, 1993.
- C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.
- C. J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297, 1984.
- W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of classification*, 20(1):025–047, 2003.
- J. Sun and C. R. Loader. Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics*, 22(3):1328–1345, 1994.
- G. R. Terrell and D. W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.
- A. B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.
- A. B. Tsybakov. Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats, 2009.
- L. Wasserman. *All of Nonparametric Statistics*. Springer-Verlag New York, Inc., 2006.
- L. Wasserman. Topological data analysis. *arXiv preprint arXiv:1609.08227*, 2016.
- B. Winter. Strong uniform consistency of integrals of density estimators. *Canadian Journal of Statistics*, 1(1-2):247–253, 1973.
- M. Woodroffe. On choosing a delta-sequence. *The Annals of Mathematical Statistics*, 41(5):1665–1671, 1970.
- Y. Xia. Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):797–811, 1998.
- J. Yukich. Laws of large numbers for classes of functions. *Journal of multivariate analysis*, 17(3):245–260, 1985.
- J. Yukich. Weak convergence of smoothed empirical processes. *Scandinavian journal of statistics*, pages 271–279, 1992.
- X.-H. Zhou and J. Harezlak. Comparison of bandwidth selection methods for kernel smoothing of roc curves.

- Statistics in medicine*, 21(14):2045–2055, 2002.
- K. H. Zou, W. Hall, and D. E. Shapiro. Smooth non-parametric receiver operating characteristic (roc) curves for continuous diagnostic tests. *Statistics in medicine*, 16(19):2143–2156, 1997.

Here is a simple illustration on how to compute the kernel density estimator (KDE) using R. We use the dataset `geyser` in the package `MASS` and focus on the variable `waiting`. We will not discuss the details of each function but only the important arguments.

## 1D case

```
library(MASS)
summary(geyser)
```

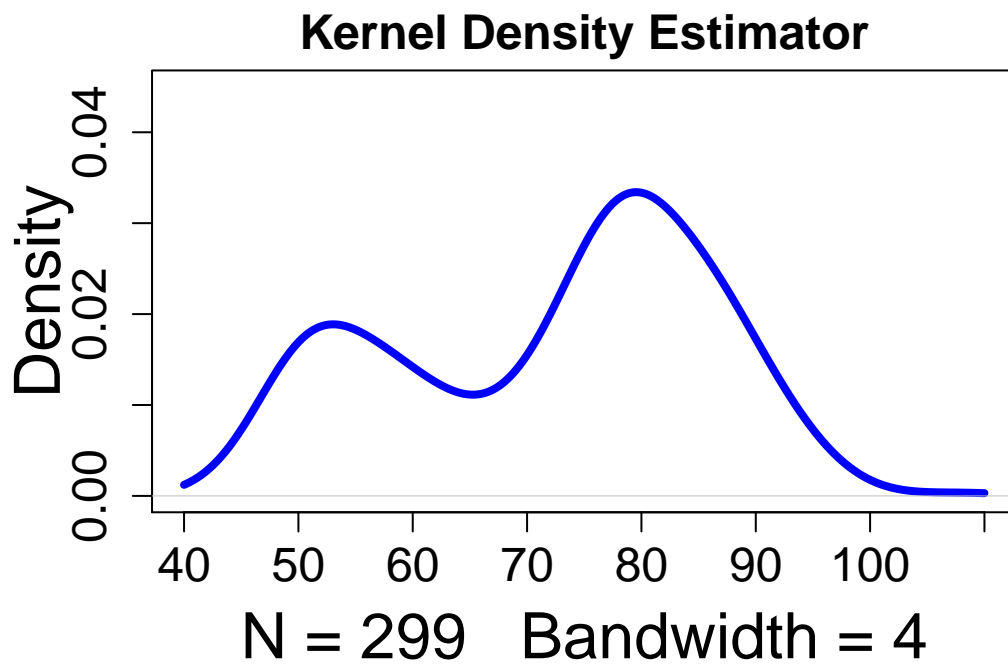
```
##      waiting      duration
## Min.   : 43.00   Min.    :0.8333
## 1st Qu.: 59.00   1st Qu.:2.0000
## Median : 76.00   Median :4.0000
## Mean   : 72.31   Mean    :3.4608
## 3rd Qu.: 83.00   3rd Qu.:4.3833
## Max.   :108.00   Max.    :5.4500
```

```
data1d = geyser[,1]
```

To use the KDE of a univariate dataset, we use the built-in function `density`:

```
h1=4
data1d_kde = density(data1d, bw=h1, from=40, to=110)
## bw = h1: we set the smoothing bandwidth = h1.
## from = 40, to = 110: the range we are evaluating the density is from 40 to 110.

##### make a plot
par(mar=c(4,4,2,1))
plot(data1d_kde, lwd=4, col="blue", ylim=c(0, 0.045), cex.axis=1.5,
      main="Kernel Density Estimator", cex.lab=2, cex.main=1.5, ylab="")
mtext("Density", side=2, line=2.2, cex=2)
```



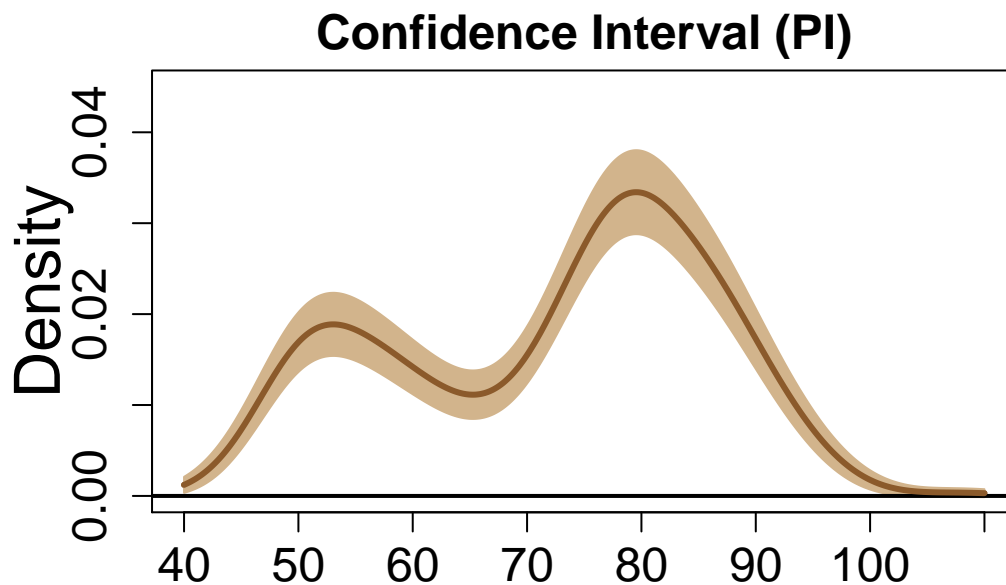
## Confidence intervals

Here we present how one may construct a confidence interval using method 1 (plug-in approach) and method 3 (bootstrap approach). Let the significance level  $\alpha = 0.05$ . Note that for the Gaussian kernel,  $\mu_K = \frac{1}{2\sqrt{\pi}}$ .

The following is how we use the plug-in approach to construct a confidence interval:

```
alpha0=0.05
n1 = length(data1d)
t0 = qnorm(1-alpha0)*sqrt(1/(2*sqrt(pi))/(n1*h1)*data1d_kde$y)
  ## data1d_kde$y: the density value evaluated at each point of data1d_kde$x.

#### make a plot
par(mar=c(4,4,2,1))
plot(data1d_kde, lwd=3, col="tan4", ylim=c(0, 0.045), cex.axis=1.5,
      main="Confidence Interval (PI)", xlab="", ylab="", cex.main=1.5)
mtext("Density", side=2, line=2.2, cex=2)
polygon(c(data1d_kde$x, rev(data1d_kde$x)),
        c(data1d_kde$y+t0, rev(data1d_kde$y-t0)),
        border="tan", col="tan")
abline(h=0, lwd=2)
lines(data1d_kde$x, data1d_kde$y, lwd=3, col="tan4")
```



For the case of bootstrap approach, here is the procedure:

```
n_BT = 1000
  ## number of bootstrap samples
kde_seq_BT_m = matrix(NA, nrow=n_BT, ncol= length(data1d_kde$y))
for(j in 1:n_BT){
  data1d_BT = data1d[sample(n1, n1, replace=T)]
  data1d_kde_BT = density(data1d_BT, bw=h1, from=40, to=110)

  kde_seq_BT_m[j,] = abs(data1d_kde_BT$y-data1d_kde$y)
}
  ## each row of 'kde_seq_BT_m' contains one bootstrap difference

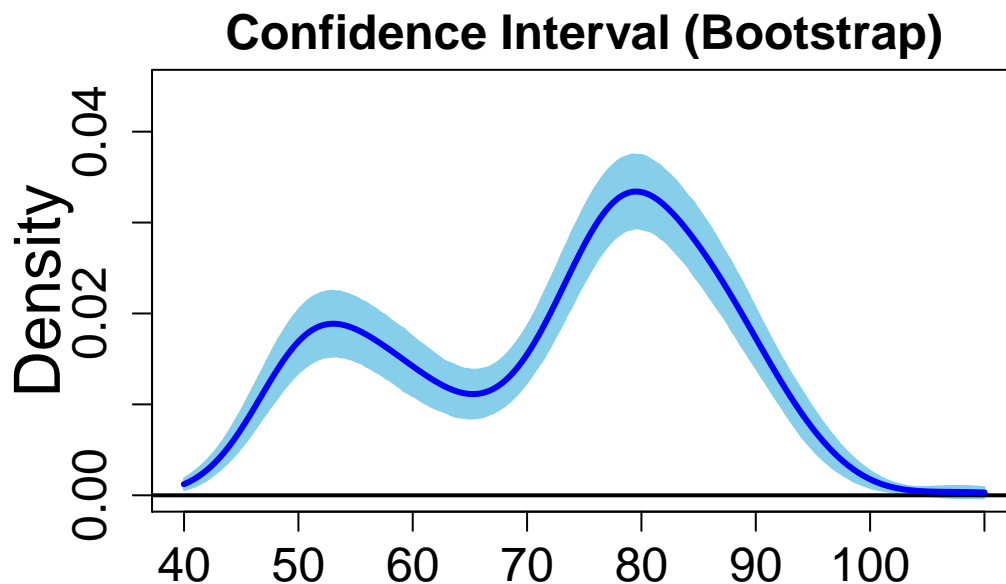
t_pt = rep(0, length(data1d_kde$y))
```

```

for(l in 1:length(data1d_kde$y)){
  t_pt[l] = quantile(kde_seq_BT_m[,l], 1-alpha0)
}
## t_pt: the 1-\alpha quantile of deviation at each point

##### make a plot
par(mar=c(4,4,2,1))
plot(data1d_kde, lwd=3, col="blue", ylim=c(0, 0.045), cex.axis=1.5,
      main="Confidence Interval (Bootstrap)", xlab="", ylab="", cex.main=1.5)
mtext("Density", side=2, line=2.2, cex=2)
polygon(c(data1d_kde$x, rev(data1d_kde$x)),
        c(data1d_kde$y+t_pt, rev(data1d_kde$y-t_pt)),
        border="skyblue", col="skyblue")
abline(h=0, lwd=2)
lines(data1d_kde$x, data1d_kde$y, lwd=3, col="blue")

```



### Confidence bands

Now we present the implementation of confidence bands. We will focus on two approaches: the traditional bootstrap approach and the bootstrapping with the debiased KDE approach, which is presented in Section 3.3.3.

The implementation of the bootstrap confidence bands is very similar to the bootstrap confidence intervals. The main difference is that we replace the pointwise difference by the uniform difference. In more details:

```

kde_seq_sup = rep(0, n_BT)
for(j in 1:n_BT){
  data1d_BT = data1d[sample(n1, n1, replace=T)]
  data1d_kde_BT = density(data1d_BT, bw=h1, from=40, to=110)
  ## bootstrap KDE
  kde_seq_sup[j] = max(abs(data1d_kde_BT$y-data1d_kde$y))
}
t_sup = quantile(kde_seq_sup, 1-alpha0)

##### make a plot

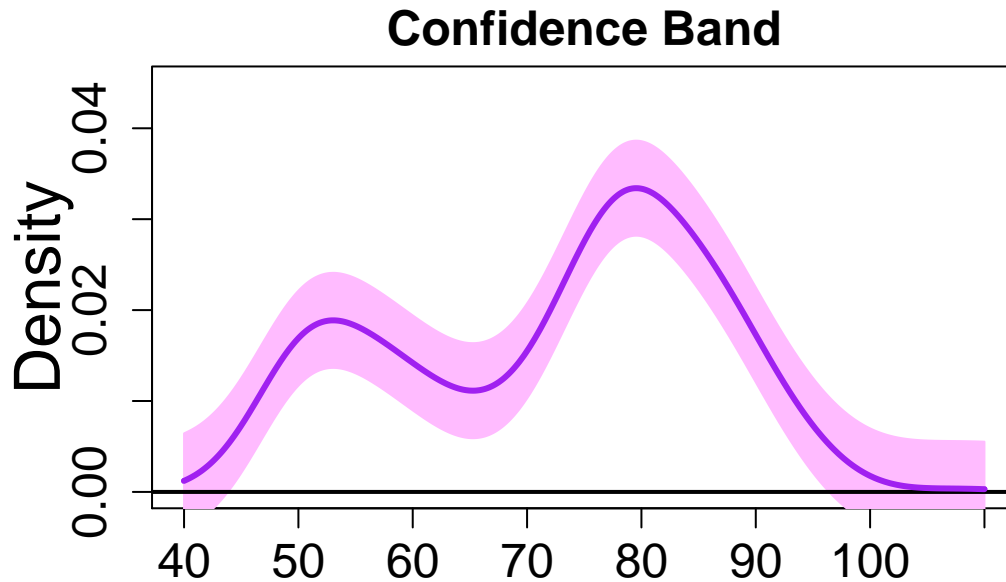
```



```

par(mar=c(4,4,2,1))
plot(data1d_kde$x, data1d_kde$y, lwd=3, col="purple", ylim=c(0, 0.045), cex.axis=1.5,
      main="Confidence Band", xlab="", ylab="", type="l", cex.main=1.5)
mtext("Density", side=2, line=2.2, cex=2)
polygon(c(data1d_kde$x, rev(data1d_kde$x)),
        c(data1d_kde$y+t_sup, rev(data1d_kde$y-t_sup)),
        border="plum1", col="plum1")
abline(h=0, lwd=2)
lines(data1d_kde$x, data1d_kde$y, lwd=3, col="purple")

```



Now we present the implementation of bootstrapping the debiased KDE approach. Recall that the debiased KDE is:

$$\tilde{p}_n(x) = \hat{p}_n(x) - \frac{h^2}{2} \cdot c_K \cdot \hat{p}_n''(x)$$

in the univariate case. The constant  $c_K = 1$  for the Gaussian kernel. Thus, we construct the confidence band by the following:

```

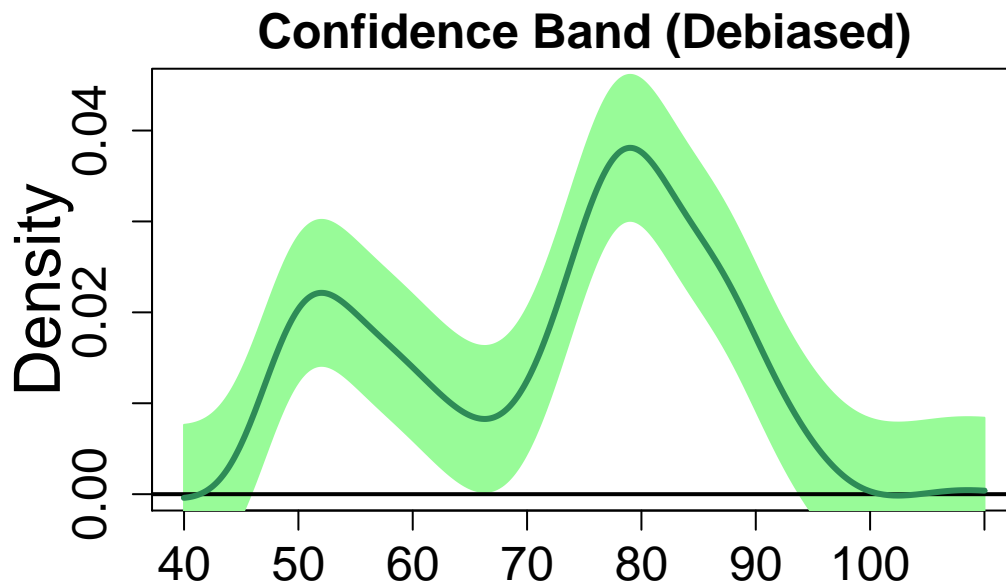
library(ks)
## this library provides density derivative estimation
X0_kdde = kdde(data1d, h=h1, eval.points = data1d_kde$x, deriv.order = 2)
## eval.points = data1d_kde$x: we evaluate the result at data1d_kde$x
## deriv.order = 2: we are computing the second derivative
de_kde_seq = data1d_kde$y - 0.5*h1^2*X0_kdde$estimate
## this is the debiased KDE, evaluated at points of data1d_kde$x

de_kde_seq_sup = rep(0, n_BT)
for(j in 1:n_BT){
  data1d_BT = data1d[sample(n1, n1, replace=T)]
  data1d_kde_BT = density(data1d_BT, bw=h1, from=40, to=110)
  data1d_kdde_BT = kdde(data1d_BT, h=h1, eval.points = data1d_kde$x, deriv.order = 2)
  de_kde_seq_BT = data1d_kde_BT$y - 0.5*h1^2*data1d_kdde_BT$estimate
  ## the bootstrap debiased KDE
  de_kde_seq_sup[j] = max(abs(de_kde_seq-de_kde_seq_BT))
}
t_de_sup = quantile(de_kde_seq_sup, 1-alpha0)

```

```
##### make a plot
par(mar=c(4,4,2,1))
plot(data1d_kde$x, de_kde_seq, lwd=3, col="seagreen", ylim=c(0, 0.045), cex.axis=1.5,
      main="Confidence Band (Debiased)", xlab="", ylab="", type="l", cex.main=1.5)
mtext("Density", side=2, line=2.2, cex=2)
polygon(c(data1d_kde$x, rev(data1d_kde$x)),
        c(de_kde_seq+t_de_sup, rev(de_kde_seq-t_de_sup)),
        border="palegreen", col="palegreen")

abline(h=0, lwd=2)
lines(data1d_kde$x, de_kde_seq, lwd=3, col="seagreen")
```

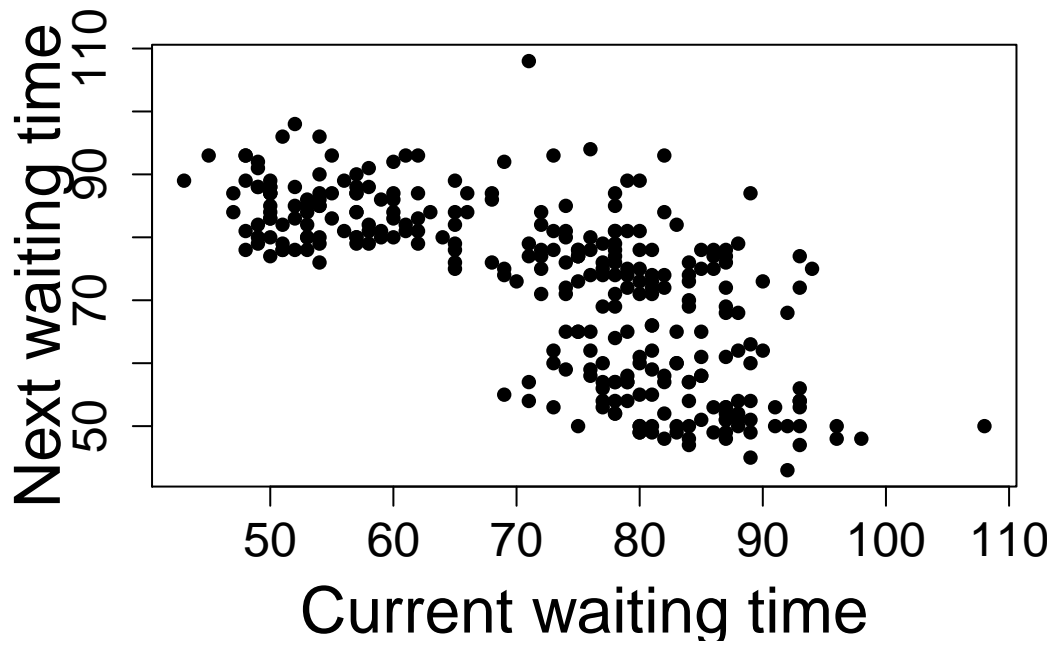


## 2D case

In the bivariate case, we recommend to use the function `bkde2D` in the package `KernSmooth`. The two variables being used are the current waiting time and the next waiting time.

```
library(KernSmooth)
data2d = cbind(data1d[1:(n1-1)], data1d[2:n1])
## getting the current waiting time and the next waiting time

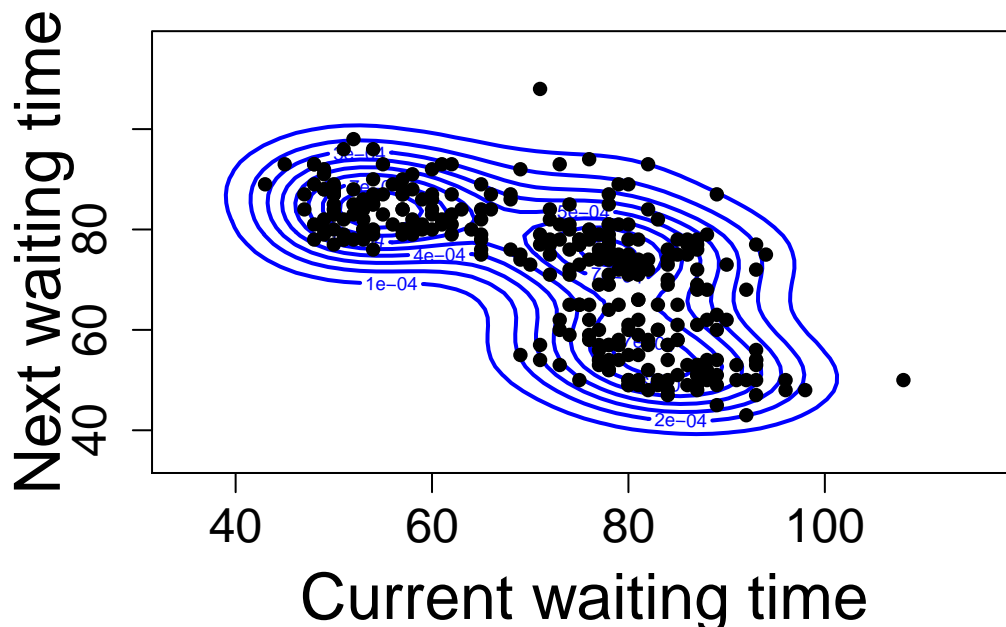
##### make a plot
par(mar=c(4,4,2,1))
plot(data2d, pch=16, xlab="Current waiting time", ylab="",
      cex.axis=1.5, cex.lab=2)
mtext("Next waiting time", side=2, line=2.2, cex=2)
```



To illustrate the 2D KDE, we use the function `contour` which displays the contour plot. We choose the smoothing bandwidth to be 5.5 (this is just an arbitrary choice).

```
h2 = 5.5
data2d_kde = bkde2D(data2d, bandwidth = h2)

##### make a plot
par(mar=c(4,4,2,1))
contour(data2d_kde$x1,data2d_kde$x2,data2d_kde$fhat, lwd=2, col="blue",
        xlab="Current waiting time", ylab="", cex.axis=1.5,
        cex.lab=2)
mtext("Next waiting time", side=2, line=2.2, cex=2)
points(data2d, pch=16)
```

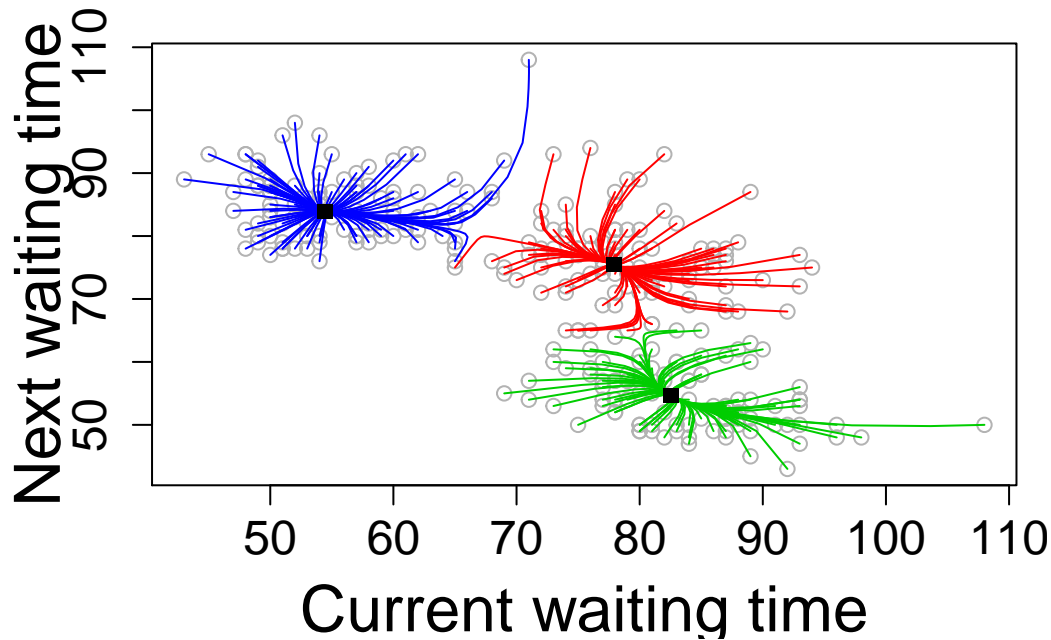


## Mode clustering

To find local modes and perform mode clustering, we use the mean shift algorithm in the package LPCM:

```
library(LPCM)

##### make a plot
par(mar=c(4,4,2,1))
data2d_ms = ms(data2d, h=h2, scaled=F, cex.axis=1.5,
               xlab="Current waiting time", ylab="", cex.lab=2)
  ## scaled = F: this will not scale the data by its range
mtext("Next waiting time", side=2, line=2.2, cex=2)
```



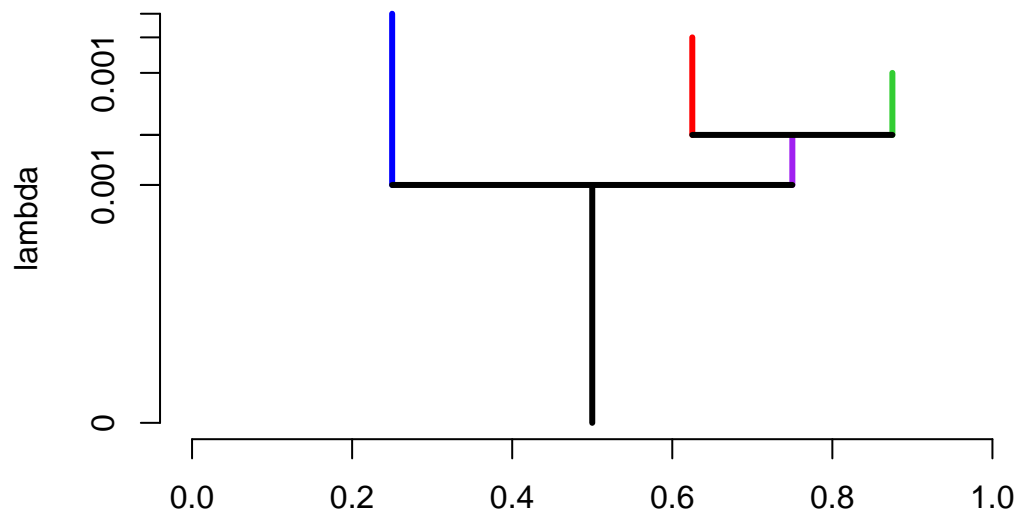
Note that the function `ms` will automatically plot the result; to disable this function, set the argument `plotms=0`. The colored curves are gradient flow lines starting at each data point. Different colors denote gradient flow lines with different destinations. The 3 destinations (local modes) are marked by black boxes. Based on the destination of the gradient flows, we partition data points into 3 clusters, denoted by the color of each flow line.

## Cluster tree and persistent diagram

Our final demonstration is the cluster tree and the persistent diagram. Both can be computed using the package TDA. For the cluster tree, we use the function `clusterTree`:

```
library(TDA)
data2d_cl = clusterTree(data2d, density="kde", h=h2, k=10)
  ## k: to compute the connected component, we need to assign this neighborhood constant.
  ## generally it is chosen to be 10-20.

##### make a plot
par(mar=c(4,4,2,1))
plot(data2d_cl, col=c("black", "blue", "purple", "red", "limegreen"))
```



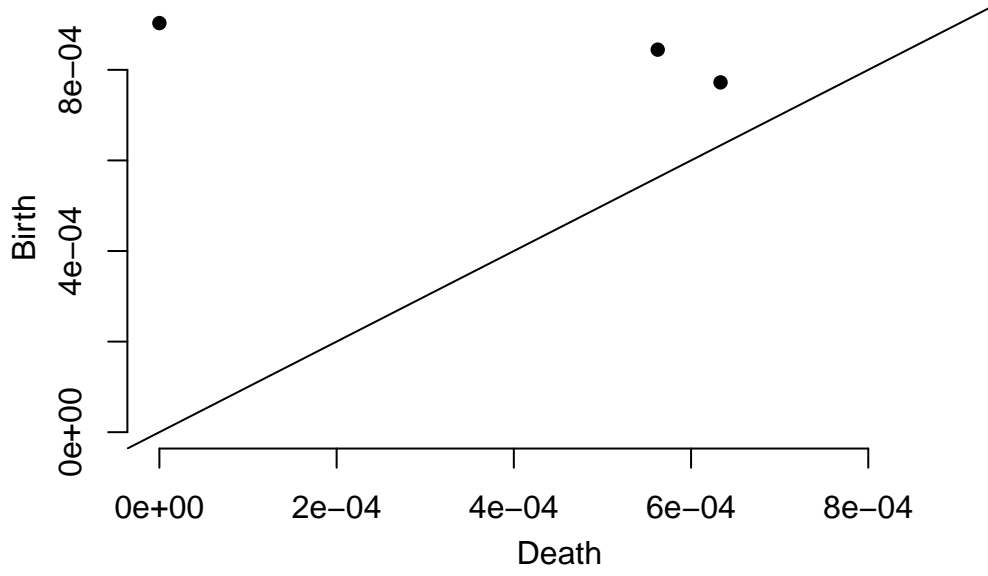
The X-axis is just an axis showing the tree structures so its numeric value does not have any meaning. The Y-axis is about the level sets at certain density levels. The top three branches are the three local modes and the corresponding connected components around them. When we move down the density threshold  $\lambda$ , every time  $\lambda$  pass the density of a local mode, a new connected component will be created. And when we keep moving down  $\lambda$ , connected components will grow, and eventually merge with others when  $\lambda$  pass certain levels (generally this will be density value of local minima or saddle points).

The purple branch appears when the connected components represented by red and green branches merged and it disappears when merging with the connected component represented by the blue branch.

Next, we demonstrate how to compute the persistent diagram of the KDE using TDA package. We will use the function `gridDiag`, which first generates a grid and then evaluate the density on the grid and compute topological features.

```
xlim0 = c(20, 150)
ylim0 = c(20, 150)
## set the lim of the grid
data2d_pd = gridDiag(data2d, kde, h=h2, lim=cbind(xlim0, ylim0), by=2, sublevel=F)
## lim = ... : the limit of the grid.
## by = 2: the separation between grid points.
## sublevel = F: the density level sets are upper (super) level sets.

##### make a plot
par(mar=c(4,4,2,1))
plot(data2d_pd$diagram)
```



The three black dots denote the birth and death time of three connected components of the KDE. When we move down the level  $\lambda$ , the birth time is the level that a connected component is created and the death time is the level that a connected component is merged into others.